Elucidating the constitutional genetic basis of multiple primary tumours

James William Whitworth

Gonville and Caius College

February 2019

This dissertation is submitted for the degree of Doctor of Philosophy

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee.

Acknowledgments

I would like to extend my sincere thanks to the following people.

To my supervisor, Eamonn Maher, for his mentorship over a time period longer than the programme of study outlined in this thesis. I am also grateful to Marc Tischkowitz, my second supervisor, for his encouragement and input.

To my colleagues past and present in the Academic Department of Medical Genetics for their insights, collaboration, advice, and camaraderie over the past few years. In particular Ruth Casey, Graeme Clark, France Docquier, Ellie Fewings, Benoit Lang-Leung, Ezequiel Martin, Eguzkine Ochoa, Faye Rodger, Phil Smith and Hannah West.

To the team at the NIHR BioResource Rare Diseases project, data from which forms the basis of this thesis. I have worked most closely with Sophie Ashford, Chris Penkett and Kathy Stirrups but there are scores of individual contributors.

To the recruitment teams at genetics centres across the UK and beyond, who make a large contribution to medical genetic research.

To the Gonville and Caius Middle Combination Room for being a great community full of interesting people and enlightening conversations.

Most of all, to my wife, Beth, for her loving enthusiasm, support and great times during the years of this PhD programme.

List of acronyms used

ACC	Adrenocortical carcinoma
ACMG	American College of Medical Genetics
AML	Acute myeloid leukaemia
AUC	Area under curve (of receiver operator characteristic)
AVL	Antoni van Leeuwenhoek (hospital)
BHD	Birt Hogg Dubé (syndrome)
CADD	Combined Annotation Dependent Depletion (score)
CATGO	Cambridge Translational Genomics Laboratory
CCDS	Consensus Coding Sequence
CNS	Central nervous system
CPG	Cancer predisposition gene
CRUK	Cancer Research United Kingdom
CT	Computerised tomography
ctDNA	Circulating tumour deoxyribonucleic acid
DSB	Double stranded break
EA	East Anglia (registry)
eQTL	Expression quantitative trait loci
ExAC	Exome aggregation consortium
FFPE	Formalin fixed paraffin embedded
GIST	Gastrointestinal stromal tumour
GO	Gene Ontology (term)
GTEx	Genotype-Tissue Expression (project)
GWAS	Genome wide association study
HGMD	Human Gene Mutation Database
HGNC	HUGO Gene Nomenclature Committee
HLRCC	Hereditary leiomyomatosis and renal cell carcinoma
IARC	International Agency for Research on Cancer
ICD-O-3	International Classification of Diseases for Oncology (3rd edition)
IGV	Integrative Genomics Viewer
IHC	Immunohistochemistry
LDL	Low density lipoprotein
LFS	Li Fraumeni syndrome
LOH	Loss of heterozygosity
MEN2	Multiple endocrine neoplasia type 2
MINAS	Multilocus inherited neoplasia alleles syndrome
MLPA	Multiple ligation-dependent probe amplification
MMR	Mismatch repair
MPMT	Multiple Primary Malignant Tumours (tag as used in NIHR Bioresource - Rare Diseases project data)
MPNT	Malignant peripheral nerve sheath tumour
MPT	Multiple primary tumours
MRI	Magnetic resonance imaging
MSI	Microsatellite instability

MTS	Multiple tumour score
NGS	Next generation sequencing
NMSC	Non-melanoma skin cancer
P/LP	Pathogenic/likely pathogenic (variant)
PAF	Population attributable fraction
PARP	Poly ADP-ribose polymerase
PCR	Plymerase chain reaction
RCC	Renal cell carcinoma
ROC	Receiver operator characteristic
SIR	Standardised incidence ratio
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SO	Sequence Ontology (term)
SPEED	Specialist Pathology Evaluating Exomes in Diagnostics (study)
SV	Structural variant
TCGA	The Cancer Genome Atlas
TCP	Tru-sight cancer panel (assay)
TSG	Tumour suppressor gene
UCNE	Ultraconserved non-coding elements (database)
UCSC	University of California Santa Cruz (genome browser)
UM	Uveal melanoma
VAF	Variant allele fraction
VCF	Variant call format (file)
VEP	Variant Effect Predictor
VHL	Von-Hippel Lindau
VUS	Variant of uncertain significance
WES	Whole exome sequencing
WGS	Whole genome sequencing
XP	Xeroderma pigmentosum

Contents

Tables and figures

Chapter 1 – Introductory chapter

This introductory chapter contains some sections adapted from text contributing to a book chapter written by the author.[1]

1.1 - Cancer as a genetic disease

The concept of cancer as a clonal expansion of cells that have undergone genomic (and/or epigenomic) changes conferring malignant properties is now broadly accepted. The development and refinement of this hypothesis has been guided by application of new technologies that have analysed cellular genetic material at increasingly higher resolution to produce previously unimagined quantities of data.

In the early twentieth century, microscopic analysis led to the observation that chromosome aberrations can occur in malignant cells.[2] Theodor Boveri made the seminal suggestion that such aberrations might be directly implicated in tumourigenesis. Studying abnormal mitoses in sea urchin embryos led him to hypothesise that disordered cellular properties, including malignancy, resulted from an unbalanced chromosome complement. Boveri proposed the existence of both "inhibiting chromosomes," i.e. those that normally act to suppress cell division and "stimulatory chromosomes," which alter a cell's relationship with its external environment to encourage a proliferative state. These ideas were prophetic of current conceptualisation of the roles of tumour suppressor genes and proto-oncogenes in the pathogenesis of human cancers.[3] Over 50 years later in the 1960's, a specific chromosomal abnormality was associated with a particular tumour when the Philadelphia chromosome (resulting from a translocation between chromosomes 9 and 22), was identified in the blood of chronic myeloid leukaemia patients.[4] Chromosomal gains, losses and rearrangements may result from genomic instability in advanced cancers but may also, as with the Philadelphia chromosome, be key to tumour initiation. With the development of DNA sequencing techniques, it became possible to study such initiating events at the individual gene level and so define causative genetic abnormalities not visible by chromosome analysis, i.e. at the nucleotide level.

1.1.1 - Oncogenes, tumour suppressor genes and cancer predisposition

The development of the concept of the oncogene was a crucial step towards understanding how genetic changes can lead to cancer. Oncogenes were initially discovered by analysing cells with malignant properties that had been induced by a retrovirus. It was found that tumorigenic potential was conferred by one component gene of the virus, which was described as the oncogene.[5] Further research revealed that orthologues of the viral oncogenes were present in normal cells, which were labelled proto-oncogenes.[5] Subsequently it was elucidated that genetic changes, unrelated to viral infection, that result in enhanced or altered function of proto-oncogenes could directly promote tumourigenesis. Proto-oncogenes are involved in a range of cellular processes that are pertinent to cell growth/proliferation including cell cycle regulation and growth signalling.

The discovery of tumour suppressor genes (TSGs), the other main gene class significant in cancer development, has been particularly relevant to the understanding of inherited tumours. Although TSG inactivation is frequent in both inherited and sporadic forms of cancers, there are many more TSGs where constitutional variants are known to cause cancer predisposition than proto-oncogenes. Under normal circumstances, a TSG often functions to inhibit cell proliferation and inherited or acquired events that induce a loss of function compromise this role, thereby promoting tumourigenesis. TSG inactivation may lead directly to cellular attributes that encourage malignant transformation or be indirect in other instances (e.g. inactivation of DNA repair genes with resultant failure to repair deleterious mutations in other TSGs or proto-oncogenes).

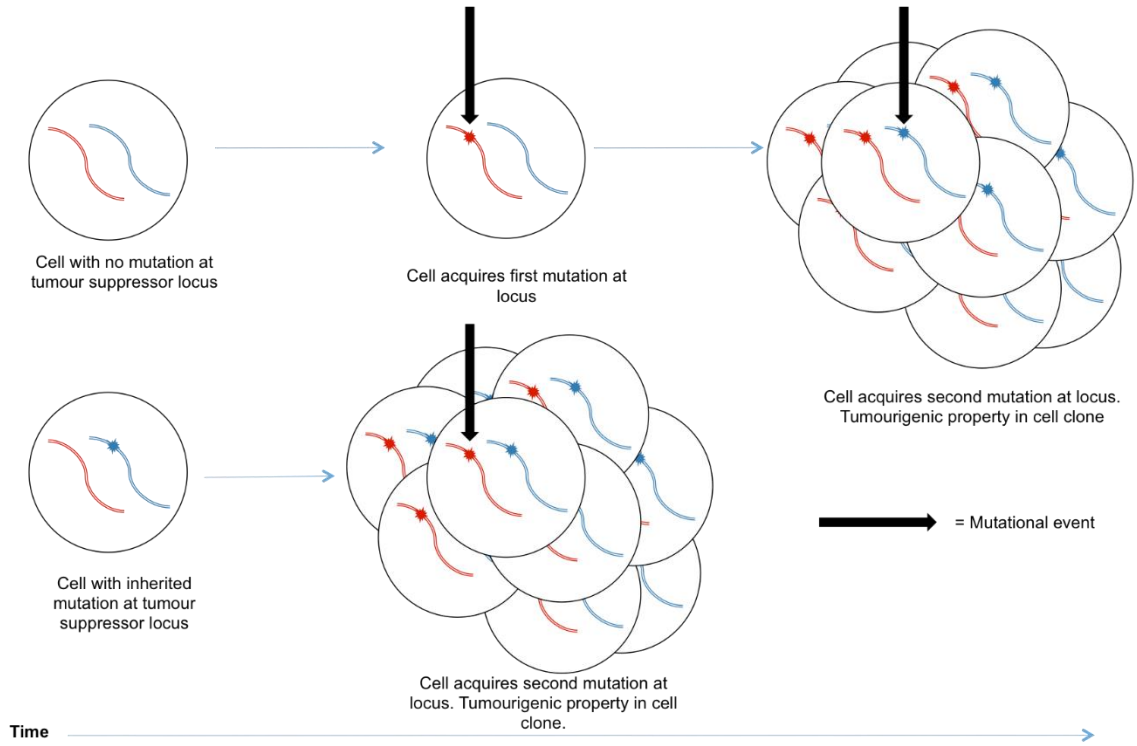
Whilst some genetic changes appear particularly important in conferring tumour defining properties to cells (such events may be referred to as driver mutations), the transition from normal cell to malignant is typically a multi-step process (though genome-wide sequencing studies have shown that the number of mutations may vary from less than 10 to thousands). In tumours that contain hundreds or thousands of mutations, normal DNA repair mechanisms are typically compromised and most of the mutations do not have a role in driving tumourigenesis (referred to as passenger mutations). A source of much debate, often based on epidemiological evidence, has been how many changes are essential for the process of tumour development. Work by Nordling observed cancer mortality correlating with age and estimated that, on average, six mutational events in a given cell were required for a cancer to occur.[6] The work only studied certain cancer types and observed that many malignancies did not conform to this model. More recently, a sequencing investigation of 29 cancer types (7664 samples) and observation of gene's non-synonymous to synonymous variant ratio (this should be high in TSGs and oncogenes) suggested around four tumourigenic mutational events are observed on average, though the number does vary between cancer types.[7]

Work by Al Knudson suggested that at in a rare embryonal tumour, retinoblastoma, the age at onset distribution was consistent with two critical rate limiting mutational events. By comparing the age at onset in familial and sporadic cases, Knudson proposed a model whereby in familial cases only a single rate-limiting mutational event ("hit") was required. These predictions were consistent with the hypothesis that in familial cases, the first rate-limiting mutation is inherited from an affected parent and that only one further hit (a somatic mutation) is required to initiate tumourigenesis. Sporadic cases, in contrast, require two somatic mutations to initiate tumourigenesis[8] (Figure 1.1). This model explains the very high risk of retinoblastoma and frequent occurrence of bilateral tumours in individuals with the familial form while sporadic cases present at an older age and have single unilateral tumours.

Familial retinoblastoma is caused by germline (constitutional) loss-of-function variants in the *RB1* TSG, which was identified through analysis of retinoblastoma tumours with Knudson's hypothesis in mind. Previous evidence existed that a region of chromosome 13 was the area undergoing the hypothesised second hit. Some individuals with retinoblastoma had been reported to harbour a constitutional deletion at this region[9] and acquired loss or partial deletion of this area of chromosome 13 had been shown in retinoblastoma tumour cells.[10] Identification of the *RB1* gene within the target region was followed by demonstration that inherited cases had an inactivating constitutional variant and a second hit in the tumour cells, whereas tumours from cases that didn't show inheritance exhibited inactivating hits of both *RB1* alleles in tumour but not normal cells (implying that both events occurred somatically).[11]

Knudson's hypothesis and the subsequent identification of the *RB1* TSG[11] was a seminal event in the development of inherited cancer genetics. Apart from highlighting the role of TSGs in cancer pathogenesis, it demonstrated that inherited constitutional variants leading to tumour predisposition could be identified through study of affected families and that genes affected by them could additionally be implicated in the more common sporadic counterparts to inherited tumours.

Figure 1.1 – Knudson's two hit model conceptual diagram[1]



The identification of *RB1* prompted a continuing search for further cancer predisposition genes (CPGs) that has yielded findings relevant to both individuals who harbour deleterious variants affecting them and those patients diagnosed with tumours occurring outside of the inherited context.

CPGs have been discovered that do not conform to a two hit TSG model and a number of constitutionally activated proto-oncogenes have been found to cause cancer predisposition (e.g. *RET* in Multiple Endocrine Neoplasia Type 2[12,13] and *MET* in hereditary papillary kidney cancer[14]). The search has focused on individuals with specific clinical characteristics of inherited cancer predisposition (e.g. young age at diagnosis of a particular cancer type) but advances in genetic technology are also providing the means for large scale sequencing in individuals with less specific features.

1.2 - The development of DNA sequencing techniques

Identification of CPGs has relied on the aforementioned development of DNA sequencing techniques that have allowed analysis of genomic regions at the nucleotide level. Around the time of, and following, the publication of the structure of DNA in 1953,[15] methods had been formulated to produce libraries of DNA fragments through techniques such as restriction enzymes and polymerase reactions.[16–19] This area would later be greatly assisted by the development of molecular cloning with recombinant DNA vectors[20] and polymerase chain reaction (PCR).[21]

A crucial advance in the analysis of libraries came with the advent of two strategies to infer the sequence of DNA by observing varying migration rates of different fragments through a polyacrylamide electrophoresis gel. Sanger and Coulson's "plus minus" technique[22] used a polymerase to produce DNA fragments of different lengths that started from the same molecular location due to the use of a single primer. Four initial reactions were undertaken where one of the four nucleotides used for extension (e.g. adenine) was radiolabelled. For each of those reactions, two further polymerisations were performed on the population of fragments containing the radiolabelled nucleotides. In one of these (the "plus" reaction), only nucleotides corresponding to the radiolabelled one (e.g. adenine) were available for polymerisation and in the counterpart reaction (the "minus" reaction), the other three were available (e.g. thymine, cytosine, guanine). When run on the electrophoresis gel, the positions of fragments (visible due to radioactivity) from the plus reaction would reveal the lengths of fragments where extension was not possible due to the fragment ending in a given nucleotide (e.g. adenine). The gel positions from the minus reaction would reveal the lengths of fragments ending in another nucleotide (e.g. thymine, cytosine or guanine). Consideration of all eight reactions could reveal the sequence of the section of DNA in question. Maxam and Gilbert[23] produced a technique with number of similarities but without using a polymerase to produce fragments of varying lengths. Instead, chemical cleavage at specific bases of radiolabelled DNA was performed and the lengths of resulting fragments from a particular cleavage reaction used to infer the positions of that nucleotide in the studied sequence. A further critical step in the advancement of sequencing was the incorporation of radiolabelled chain terminating nucleotides into the polymerase reactions of Sanger's technique that did not have a 3' hydroxyl group necessary for extension of the

nucleic acid sequence.[24] If four polymerase reactions were performed where a proportion of the nucleotide pool is made up of a single type of chain terminating nucleotides (e.g. adenine), a population of fragments ending in that base would be produced. The relative positions of fragments from the four reactions on an electrophoresis gel would subsequently reveal the template sequence. Sanger sequencing was developed further by the substitution of radiolabelling for fluorescent chain terminating nucleotides, allowing a single polymerase reaction as they could be visually distinguished from each other. The electrophoresis gel was also substituted for capillary electrophoresis where the chain terminating nucleotide colour detected by a fixed camera at a given time could be used to infer the last base of a particular fragment size.[16]

Sanger sequencing based techniques formed the basis of most sequencing performed in the latter part of the twentieth century, including that contributing to the human genome reference sequence published in 2001.[25] Although improvements efficiency had taken place, these processes remained reliant on separate reactions to sequence each template fragment of interest, which were limited in length. The parallelisation of reactions increased the scope of sequencing significantly and brought about the techniques widely referred to as next generation sequencing (NGS). An early NGS method was developed by 454 Life Sciences,[26] which incorporates synthetic adapter sequences to a potentially large library of DNA molecules, allowing attachment to beads (optimally one molecule per bead). A PCR reaction is then used to amplify the DNA attached to each bead ready for separate sequencing reactions. Crucially, each bead is attached to a fixed position on a solid surface from which sequencing readout pertaining to that bead will be measured. Rather than utilising chain terminating nucleotides, sequencing reactions proceed by pyrosequencing. Pyrosequencing still produces sequence readouts through synthesis based on a template but measures the release of pyrophosphate that occurs when a nucleotide is incorporated into a growing complementary DNA strand in real time. Pyrophosphate can be converted to adenosine triphosphate, which in turn can form the substrate for a fluorescent luciferase reaction. The relevant enzymes are introduced to the solid surface along with a nucleotide pool of a single type. If the next position in the growing DNA strand is complementary to that base it will be incorporated and light emitted. Reactants are subsequently washed away and the process repeated for the other three nucleotides.

The NGS platform that would become mostly widely used was developed by Solexa and later acquired by Illumina, whose products provided the sequencing data for this project. This technique[27] still utilises adapter sequences and location onto a solid surface (referred to as a flow cell) but hybridises adaptors to complementary oligonucleotides rather than beads. Fixed molecules subsequently undergo PCR amplifications at their respective locations. Illumina NGS sequencing reactions use a pool of all four fluorescent chain terminating nucleotides during polymerisation. Incorporation of a particular nucleotide in the growing synthesised DNA molecule and washing away

of the other nucleotides allows light emission of a colour corresponding to the incorporated nucleotide, which is detected at the relevant position on the flow cell. In contrast to Sanger sequencing, the fluorescent and chain terminating portions of the incorporated nucleotide are then chemically removed and incorporation of nucleotides can continue along the growing molecule. An advantage of this approach is that it avoids inaccuracies associated with pyrosequencing of homopolymer tracts as only one nucleotide is incorporated at a time. A limitation is the relatively short length of sequence readouts (reads) that can be obtained, which has implications for accuracy of alignment to reference sequences. However, the technique produces reads from each end of a DNA template (paired end data) that can be used to make inferences such as whether a deletion exists in a genomic region (indicated by a longer than expected insert size between two paired reads).

Whilst Illumina products remain the dominant sequencing platforms, other technological developments have led to further advances and a group of assays referred to as third generation sequencing. These techniques are characterised by the sequencing of single (rather than amplified) molecules and the production of long reads that facilitate alignment to (or production of) reference genomes. They are particularly relevant to some applications such as identifying structural variants due to an increased chance of reads being generated containing a chromosomal breakpoint. Prominent platforms include those produced by Pacific Biosystems[28,29] and Oxford Nanopore.[30] The former utilises nano-engineered wells that are small enough to induce rapid decay of light from a laser source as it penetrates the well. Consequently, a visualisation area at the base is created where a single DNA polymerase is bound. Exposure of wells to template DNA molecules and fluorescently labelled nucleotides induces strand synthesis where incorporation of specific nucleotides can be visualised through their fluorescent signal with minimal noise due to the small size of the visualisation area. Rather than observing nucleotide incorporation by DNA polymerase, Oxford Nanopore technology passes single stranded DNA molecules through a nano-pore following denaturation by an enzyme located at the pore entrance. The pore is embedded in a polarised membrane, inducing movement of the DNA through it and creating ionic flows that are altered in a characteristic manner by the passage of particular bases. The sequence of bases can be inferred from measurement of these. One source of excitement with this platform is its miniaturisation that makes use at the bedside or in the field feasible.

NGS applications produce a series of reads from the sequenced population of molecules with no accompanying information regarding genomic location from which they were sequenced or whether they show any evidence of variation from a reference sequence. NGS sequencing data can be translated into variants as a result of extensive work that has taken place to produce software and algorithms for this purpose. A key initial step is the alignment of each read to its corresponding position in the chosen reference sequence. This can be performed with a number of tools but this

project utilised the widely used Burrows Wheeler Aligner[31] and more recently developed Illumina Isaac.[32] In an ideal situation, each read is confidently aligned to a unique genomic position but similarities between different regions can produce multiple alignments for the same read. This begets uncertainties as to whether apparent sequence variation in the read results from genuine deviation from the reference or an origin from a similar but different location. Alignment can then be followed by variant calling whereby the most likely sequence at a given site is calculated on the basis of a number of lines of evidence such as quality of the base call in the read (incorporated pre-alignment), the number of reads supporting a particular base call and the extent to which base calls in the read match those in the reference genomic region it is aligned to. Like alignment, a number of different variant calling tools exist with Genome Alignment Toolkit HaplotypeCaller[33] (preceded by Unified Genotyper) being the most widely used for germline variants and Illumina Isaac also being used in this work.

The descriptions of NGS techniques above emphasise commonalities between workflows but there are key variables in the processes that influence suitability for the question and resources at hand. A key parameter is the genomic regions covered by sequencing reads, which may range from a single gene to close to an entire genome (whole genome sequencing). Techniques that selectively produce sequencing reads aligning to a number of pre-defined genes are frequently referred to as gene panels, whereas incorporation of all coding regions can be designated exome sequencing. The pre-sequencing DNA library preparation steps for these outcomes are frequently similar and generally involve the hybridisation of oligonucleotide sequences corresponding to regions of interest to molecules in the library. The oligonucleotides are designed with modifications (e.g. magnetic beads or biotin) to enable their physical extraction along with the library molecules they are attached to, a process referred to as capture. Panel or exome sequencing also usually involves a PCR amplification step where primers attach to common adaptor sequences introduced to library molecules. This might be prior to capture where all fragmented molecules can be amplified, or afterwards when only molecules of interest undergo PCR. Whole genome sequencing does not require capture or PCR and libraries prepared for it are the result of fragmentation of a sample with subsequent ligation of adapter sequences at random. Any genomic region is eligible for coverage by sequencing reads and the lack of a PCR step also reduces variation in coverage as regions that are difficult to amplify are less likely to be under-represented.

A further important distinguishing feature between NGS applications is sequencing depth, which refers to the number of reads that align to a given base and is expressed as 1X, 2X and so forth. A higher number of reads produces greater confidence of a variant call, particularly if somatic variants are sought. However, higher depth may be associated with greater resource expenditure and high confidence calls can still be made with lower numbers of reads. The number of generated reads can be

influenced by numerous variables in the sequencing process (e.g. starting DNA quantity, number of samples per flow cell used) but depth is frequently inversely correlated with extent of genomic coverage. Gene panels often provide hundreds to thousands of reads per target base whereas whole genome sequencing depth is typically well below 100X.

An additional key variable in NGS sequencing is read length. Much of the recent development in genomics has been based on short read sequencing, typically in the region of 100-150bp. However, long read techniques such as those mentioned above provide the opportunity to increase this number to multiple thousands. Advantages include more specific alignment to reference sequences with consequent reduction in reads mapping to multiple sites and spurious base calls. Structural variant calling can also be improved due to the greater chance of observing reads crossing chromosomal breakpoints and phasing of variants is facilitated by an increased probability of individual reads covering areas in which multiple variants lie.

1.3 - Identifying cancer predisposition genes

A variety of different study designs have been used to identify CPGs, generally using one of the sequencing techniques described above in combination with a strategy to narrow down the genomic region of interest.

Earlier efforts focused on large families with multiple affected members and used genetic linkage to elucidate regions that segregated with cancer incidence. This strategy is greatly assisted by high penetrance of variants affecting the CPG that is sought. In some cases (e.g. *RB1*) the suggested CPG location was supported by the identification of deletions/allele loss in tumour material. Having defined a region containing the putative CPG that was as small as possible, all genes within the region were then sequenced to identify those that were recurrently mutated. Some CPGs (e.g. *APC* and *VHL*) are frequently somatically mutated in cancer and these observations can assist in proposing regions or genes as candidates. Occurrence of multiple constitutional variants in a given gene amongst individuals with the cancer in which the somatic mutations were reported can be taken as evidence of its status as a CPG. Recently, molecular characterisation of tumours has accelerated and is assisting with investigation in this area. cBioPortal, for example, contains data from over 70,000 sampled tumours from a variety of cancer genome sequencing projects.[34]

Previously, technological and resource constraints meant that genes/regions to be sequenced were highly targeted. NGS platforms have enabled analysis of whole genomes, coding regions (exomes) or a selected series of genes at a cost that is realistic for many research groups. NGS has greatly facilitated CPG identification in projects that often start with a less defined hypothesis in terms of a candidate gene/region. The challenges presented by the resultant large numbers of rare genetic

variants are often significant but combination with other lines of evidence can filter causative candidates. Evidence can include gene expression in the tissue of interest or involvement of the candidate gene product in a cellular process relevant to cancer (e.g. DNA repair). Segregation of variants within families with multiple affected members remains critical in many analyses and it is notable that in the study reporting *POLE* as a CPG described below, an initial search for shared coding variants between unrelated probands did not produce any firm candidate genes.[35]

Recent examples of CPG discovery using NGS techniques include studies of individuals with colorectal adenomas and cancer that causally implicated *NTHL1*[36] and *POLE*.[35] The former study applied exome sequencing to 51 individuals (from 48 families) with multiple colonic adenomas and focused on truncating variants shared between unrelated participants. Under a recessive inheritance hypothesis, five genes were identified that contained such variants, one of which (*NTHL1*) was a DNA repair gene. Four individuals from three families had a biallelic truncating variant in this gene that was not detected in controls. *POLE* was identified as causing adenomas through whole genome sequencing of members of a single family with multiple affected individuals. The analysis took advantage of pre-existing linkage analysis in the family to restrict the area of investigation to a small number of genomic regions. Six non-synonymous coding variants in those regions were shared between all three sequenced cases, one of which was within a gene (*POLE*) with relevance to DNA repair as it encodes a protein product with a polymerase proof reading function. The putative causative variant was then identified in 12 out of 3805 additional colorectal cancer cases used as a validation set and no controls.

A further approach that has yielded success in CPG identification is that of a case control analysis whereby frequency of variants in a given gene is compared with that in a set of controls. If deleterious variants in a CPG confer moderate cancer risks, multiple variant carriers in a kindred are likely to be unaffected due to incomplete penetrance. Therefore, segregation data to narrow down candidate variants may be misleading. Case control studies do not rely on multi-case families but are greatly assisted by large numbers of participating individuals. A number of CPGs have been identified by undertaking sequencing in breast cancer cohorts, a common tumour type facilitating a high number of participants. *CHEK2* was proposed as a CPG due to its role in DNA repair and interaction with *BRCA1*. A founder truncating variant was found to be significantly more frequent in breast cancer cases vs controls and estimated to lead to a doubling of risk.[37] Similarly, *PALB2* associates with *BRCA2*, a line of evidence that helped identify it as a breast CPG in a study observing truncating variants in 10 out of 923 familial breast cancer probands compared with zero out of 1084 controls.[22]

1.4 - Risks associated with variants in cancer predisposition genes

NGS technologies have assisted novel CPG discovery but pathogenic variants affecting many of them are often estimated to cause lower tumour risks than some earlier discoveries such as *APC* and *TP53*. Most CPGs that affect large numbers of individuals, and in which high penetrance variants occur may have been discovered.

Newly identified high risk CPGs are likely to be rare and consequently account for a very small proportion of population cancer burden. Despite this, associated clinical utility will be significant for affected individuals and can provide insights into similar tumours that are not due to constitutional variants in the CPG in question. Furthermore, any contribution to a greater range of variants known to be relevant to a particular cancer phenotype can be incorporated into a more comprehensive diagnostic test. Such a test has an enhanced negative predictive value in patients who consult with the relevant phenotype but, as is often the case, do not receive a molecular diagnosis explaining their tumours.

Case control based analyses can reveal significant association of variant/gene with tumour without necessarily reflecting a very high risk of that neoplasm developing. *BRIP1* and *PALB2,* for example, were originally reported to confer a relative breast cancer risk of 2 and 2.3 respectively.[38,39] Interestingly, further observations of variant carriers has revised the *PALB2* associated risk to a much higher level[40] and refuted the role of *BRIP1* truncating variants in predisposing to breast cancer,[41] illustrating that risks associated with CPG variants are far from static. One factor contributing to this flux can be the precise variant composition of studied cohorts. A large multi-centre analysis involving 42,671 breast cancer cases and an equal number of controls noted variant frequency and estimated risks for ten rare variants in *PALB2, CHEK2* and *ATM*. Risks were often comparable to those earlier gene level based estimates but varied substantially between variants in the same gene.[42]

Elucidation of the genetic basis of high to moderate penetrance cancer predisposition phenotypes can have a large effect on management of affected families but only impact on a small minority of cancer patients or at-risk individuals. Genome wide association studies (GWAS) are case control studies of large cohorts of cancer patients that reveal more common genetic variants associated with a small increased risk of particular cancers in larger numbers of individuals. Identification of alleles such as these can provide insights regarding the molecular pathways significant to particular tumours but generally haven't been translated into preventative clinical settings because the associated increased risk is not sufficient to prompt specific interventions such as surveillance imaging. However, clinical utility might be provided by identifying individuals with multiple risk alleles that may act in combination. One report to assess the potential of this approach analysed risks associated with combinations of 77 variants which had been previously associated with breast cancer in GWAS

studies. A combined polygenic risk score was formulated, which was used to stratify over 30,000 breast cancer cases and controls into risk quintiles. In those without a family history, the upper quintile had a higher lifetime risk of breast cancer (16.6%) than the lower quintile (5.2%). This difference was more pronounced in those with a first degree relative with breast cancer (24.4% vs 8.6%).[43] Risk estimates in the top quintile group, therefore, approach those deemed sufficient for risk mitigating intervention in the clinic. A more recent report applied polygenic risk scores based on 18 breast cancer associated GWAS variants to 9222 women seen in a breast cancer family history clinic and observed a twofold difference in risk between the top and bottom score quintiles (in women without *BRCA1*/*BRCA2* pathogenic variants).[44]

1.5 - Cancer predisposition genes and their contribution to cancer burden

The canon of CPGs has reached three figures in number but defining such a gene is not without difficulty. A comprehensive review of CPGs was published in 2014 by Rahman and included genes where rare pathogenic variants at least double the relative risk of a given cancer type and lead to at least 5% of carriers being affected with cancer.[45] For some tumours, it is doubtful whether the lower end of these risks would be of benefit for clinical management.

The proportion of cancers attributable to inherited cancer syndromes is not easily arrived at due to the lack of a clear definition of a CPG and limited information regarding frequency of pathogenic variants in the population or the precise tumour risks conferred by them. The figure is often estimated as around 10%[46] and has been quoted as 3% if only known CPGs are included in the estimate.[45] A recent analysis of germline sequencing data from participants in The Cancer Genome Atlas (TCGA) database searched for rare protein truncating variants in 114 CPGs and identified them in differing proportions between cancer types. Figures ranged from 4% in acute myeloid leukaemia and glioblastoma to 19% in ovarian cancer.[47] Overall, the proportion of cases with CPG truncations was roughly consistent with the 10% estimate previously proposed but other types of variant such as missense are known to contribute to cancer predisposition and undiscovered CPGs not on the list of 114 may also be significant.

An estimate of the contribution of all genetic factors to cancers can be arrived at by assessing heritability, which is the estimated proportion of variation of a trait (in this case liability to develop a tumour) in a population that is accounted for by genetic factors and not by environmental factors or chance. Heritability estimates can be derived from observing the incidence of a given cancer type amongst relatives of individuals who develop that malignancy and comparing it with incidence in the general population from which they were drawn. Any excess incidence is likely to be due to genetic commonalities. Concordance of occurrence of a wide range of common cancer types in monozygotic vs dizygotic twins have been examined in Scandinavian population-based registries. An advantage to

this approach is the ability to distinguish genetic from shared environmental factors as monozygotic twins have greater genetic commonality than dizygotic twins but there is likely to be little difference between the two twin types in terms of environmental exposure. Estimated heritability ranged from 27% (breast) to 42% (prostate).[48] A later analysis of twins in the same registry was able to include 80,309 monozygotic and 123,382 dizygotic twins. The overall heritability of cancer was estimated at 33% with estimates for breast and prostate remaining similar at 31% and 57% respectively. The highest heritability estimate was 58% for melanoma, illustrating that variability in risk largely explained by genetic factors does not imply that environmental factors (in this case ultraviolet light exposure) are insignificant in the individuals developing a given cancer type.[49] Analysis of cancer cases in the Swedish Family-Cancer Database estimated genetic contribution through comparison of incidence in closer vs more distant relatives and provided estimates of between 1% and 53% depending on cancer type with thyroid cancer being the highest.[50] The heritability estimates described above relied on comprehensive registration of twins, cancer occurrences and familial relationships between individuals contained in cancer registries. Such resources are not widespread, hampering efforts to include greater numbers of individuals and apply estimates to more population groups. Even in comprehensive twin registries, rarer cancers may not be frequent enough to derive accurate estimates.

Disparity between estimates of proportion of cancers due to recognised predisposition syndromes and total heritability suggests that ongoing efforts to identify individuals with constitutional genetic factors leading to neoplasia in research and clinical settings may be rewarding. The architecture of such factors is likely to be diverse in terms of number of loci involved in a given individual and in the nature of mutational mechanisms. Heritability estimates do not give a strong indication of whether increased tumour incidence in more closely related individuals is due to a combination of numerous lower penetrance alleles or a single high penetrance CPG variant. Constitutional single nucleotide variants and indels in coding regions may account for a proportion of unrecognised predisposition syndromes but other less readily detectable mechanisms are also likely to be significant in many cases. These include structural variants, somatic mosaicism, epimutation and variation in non-coding regions.

1.6 - Mendelian conditions due to variants in cancer predisposition genes

CPG functions are relevant to a variety of cellular processes where disrupted function can beget tumourigenic phenomena such as abnormal cell cycle regulation, genomic instability or proliferation. Tumour predisposition conferred by CPG variants can often produce sufficient risks for Mendelian inheritance patterns to be observed in affected families but unaffected variant carriers may still exist in these kindreds due to incomplete penetrance. Most such conditions that have been described show an autosomal dominant inheritance pattern where bi-allelic pathogenic variants may be embryonically

lethal (e.g. *BRCA1*). A smaller number of recessive syndromes are also known including colorectal polyps and cancers due to bi-allelic pathogenic *MUTYH* variants.[51] Interestingly, a number of CPGs are associated with distinct phenotypic effects depending on whether deleterious alleles are present in the mono-allelic or bi-allelic state. Pathogenic *SDHB* variants are associated with phaeochromocytoma and paraganglioma in the heterozygous state[52] whereas bi-allelic inheritance causes a neurodevelopmental disorder.[53] Homozygous or compound heterozygous *ATM* deleterious variants were previously identified as the cause of Ataxia Telangiectasia, a childhood onset condition causing a number of features including cerebellar ataxia, immunodeficiency and predisposition to haematological malignancies.[54] The observation of increased breast cancer incidence in heterozygous carriers[55] helped to define mono-allelic variants in *ATM* as causative of a moderate risk of that tumour. In situations where there is a contrasting phenotype between mono and bi-allelic CPG variant carriers, it is possible that tumour risks associated with the mono-allelic state are still present where two deleterious alleles are inherited but that these manifestations are infrequently observed due to the recessive condition reducing life expectancy. Indeed, some occurrences of breast cancer have been reported in individuals with Ataxia Telangiectasia surviving for a longer period.[54]

1.6.1 - Tumour spectrum associated with cancer predisposition genes

Collectively, cancer predisposition syndromes can increase the risk of a large number of topographical and morphological tumour subtypes. Some inherited cancer syndromes, such as Li-Fraumeni syndrome, are associated with an increased risk of a wide range of cancer types but most conditions are currently known to predispose to a smaller number of specific tumours. Even Li-Fraumeni syndrome related cancers are among a set of four core malignancy types in 70% of cases.[56] The reason for this specificity is largely yet to be elucidated. Theoretical explanations include aberrant cellular mechanisms rendering cells susceptible to further mutation through organ specific environmental exposures (e.g. skin exposure to ultraviolet light in Xeroderma Pigmentosum) and relative functional redundancy of the relevant CPG in low risk tissues. One intriguing possible mechanism for the latter is compensation through expression of CPG paralogues derived from the same ancestral gene. A recent study of disease gene paralogue expression across multiple tissues showed that lower levels of expression are observed in tissues that are affected by variants in corresponding disease genes, but the report was primarily concerned with non-CPGs.[57]

Some phenotypic specificity may be explained by ascertainment biases influencing the study of CPGs and their associated tumour spectra. Identification of CPGs has generally occurred by preferentially studying families where there are multiple occurrences of the same tumour type, restricting other possible associations. The identification of novel CPGs in these scenarios is likely to underestimate the range of tumours caused by variants in that gene. These effects may be exacerbated by the effect of clinical criteria used to guide access to genetic testing which further extend ascertainment bias.

Widening of the phenotype associated with a CPG after initial identification based on a single cancer type is exhibited by the relatively recently described *BAP1*. This gene was originally reported as a CPG through the study of uveal melanoma (UM) cases. Previous evidence existed for a role of *BAP1* in the tumourigenesis of UM such as the observations that it is somatically mutated in around half of UM's[58] and is located on chromosome 3, which is often deleted in UMs.[59] Germline sequencing of 53 UM probands with clinical evidence of inherited predisposition showed a truncating *BAP1* variant in one individual, whose tumour demonstrated loss of the wild type allele and reduced immunohistochemistry staining for the protein product. This pattern was also found in a lung adenocarcinoma from the individual as well as a meningioma from a relative who also carried the variant.[60] Subsequently, constitutional *BAP1* variants have been associated with a range of other tumours including renal cell carcinoma (RCC) where a segregating splice site variant was found in a family with four affected individuals. Further analysis of 60 families with clustering of RCC and other *BAP1* related tumours showed variants in 11.[61]

RCC has also been observed as an additional phenotype associated with constitutional *SDHB* variants, which were initially identified as predisposing to phaeochromocytoma and paraganglioma through sequencing of the gene in affected kindreds. Study of *SDHB* was prompted by prior knowledge of a gene encoding another succinate dehydrogenase enzyme subunit (*SDHD*) causing similar phentoypes.[52] Subsequently, RCC was observed in two families with *SDHB* related paraganglioma with loss of heterozygosity shown in all of the kidney tumours.[62] Prompted by this and the rationale that *FH* variants can cause RCC and are within a gene that encodes another Krebs cycle enzyme, *SDHB* was sequenced in 68 individuals with familial and/or early onset RCC with variants identified in three.[63]

Rare cancer predisposition syndromes that become established in clinical practice may accumulate novel tumour associations through the development of larger series of affected individuals, often contributed to by multiple centres. Pathogenic variants in *PTEN* cause a range of disorders including Cowden syndrome, which is characterised by macrocephaly, cutaneous manifestations and cancer of the breast, thyroid and endometrium. However, a study of 368 *PTEN* variant carriers showed increased standardised incidence ratios (comparison of adjusted incidence vs general population) for colorectal cancer, renal cell carcinoma and melanoma.[64] These newly documented associations were arguably made possible by collaborative efforts to collect sufficient numbers of cases with the intention of better defining phenotypes caused by *PTEN* variants.

1.6.2 - Penetrance of cancer predisposition gene variants

Whilst elucidating the full tumour spectrum associated with a CPG is of critical importance, clinical utility is also derived from accurate penetrance estimates regarding the tumours affected individuals are known to be at risk of developing. Accuracy is assisted by the observation of large numbers of cases, making estimates more difficult for rarer cancer predisposition syndromes. Even where relatively large numbers of cases are diagnosed, risk estimates can be influenced by a range of factors.

Ascertainment biases can influence estimated penetrance as well as associated tumour spectrum because individuals where the phenotype is more severe e.g. earlier age of tumour diagnosis, may be prioritised for clinical testing. Studies of known variant carriers may consequently over-estimate risks, which appears to have occurred in research surrounding Lynch syndrome. Lynch syndrome increases the risk of colorectal cancer and is caused by heterozygous variants in mismatch repair genes including *MLH1, MSH2* and *MSH6*. Colorectal cancer has a high population frequency and criteria have previously been developed to prioritise clinical testing and/or research for families likely to be exhibiting tumours caused by Lynch syndrome rather than another cause. The two primary examples are the Amsterdam criteria,[65] which require a prominent family history for fulfilment and the Bethesda criteria,[66] which were developed to guide molecular investigation for suspected Lynch syndrome and allow for the inclusion of a greater number of families whilst still requiring reasonably strong evidence. Use of such criteria to ration molecular investigation can lead to efficient use of resources but may over-estimate the tumour risks conferred by deleterious mismatch repair gene variants because families with lower risks (perhaps due to a different pattern of modifying genetic variants) are less likely to have been eligible for testing. Risk of colorectal cancer due to Lynch syndrome has reduced with more recent studies compared with those conducted at an earlier time point when molecular analysis was more restricted. A large registry based analysis of Finnish pathogenic mismatch repair gene variant carriers in 1999 reported a cumulative colorectal cancer incidence of 82% by age 70.[67] However, an assessment ten years later based on carriers identified through genetics clinics and corrected for ascertainment bias estimated an equivalent figure of 66%.[68] Ascertainment biases can be reduced by prospective observation of cancer incidence in CPG variant carriers and a more recent study recorded this in 1,942 carriers of pathogenic variants in Lynch syndrome genes.[69] Cumulative incidence of colorectal cancer was lower still and reported as 46% for *MLH1*, 35% for *MSH2*, 20% for *MSH6* and 10% for *PMS2*. Risk estimates are not uniformly reduced to this extent through the application of prospective observations. Hereditary Breast and Ovarian Cancer is a further cancer predisposition syndrome where accumulation of large cohorts of pathogenic variant carriers has occurred and a retrospective meta-analysis of studies in 2003 incorporating 289 *BRCA1* carriers estimated a cumulative breast cancer risk of 65% by age 70 years.[70] A collaborative prospective analysis 14 years later included 6,036 carriers and estimated a similar risk of 72% by age 80 years.[71]

Studies of Hereditary Breast and Ovarian Cancer have highlighted a further influence on risk estimates, that of family history. Extent of family history is frequently taken as a proxy measure for genetic modifying factors that influence cancer risk in addition to the pathogenic CPG variant in question. The aforementioned prospective analysis stratified cumulative cancer risks according to family history status. For example, *BRCA1* pathogenic variant carriers with no family history of breast cancer had a 53% cumulative risk of breast cancer by age 70 years but those with at least one first or second degree relative diagnosed with that tumour type had a cumulative risk of 71% by the same age.[71]

Identification of CPG variants through clinical testing prompts assessment of pathogenicity using various lines of molecular, clinical and literature-based evidence. If the conclusion from the diagnostic service is that the variant is pathogenic, patients are frequently managed according to risk estimates that are the same for all or most pathogenic variants affecting the gene in question. However, the phenotypic effects of different pathogenic variants in the same gene can be contrasting. Multiple Endocrine Neoplasia Type 2 (MEN2) is caused by activating missense variants in the *RET* proto-oncogene and is associated with a range of neoplasms including parathyroid hyperplasia/adenoma, medullary thyroid cancer and phaeochromocytoma.[72] The chance of developing these tumours is known to be influenced by the codon in which the variant occurs and specific genotype is incorporated into clinical management guidelines. Codon 634 variants lead to higher risk of phaeochromocytoma that prompts biochemical screening from eight years of age as opposed to 20 years as per many other variants.[73] In addition, cutaneous lichen amyloidosis is observed, which is not seen in carriers of variants in other codons.[74] Met918Thr is only known to cause the MEN2B clinical subtype, which is associated with additional manifestations such as gastrointestinal ganglioneuromatosis.[75] Some variant consequences such as premature stop codons are frequently taken as indicating a complete loss of function of the affected allele but there is variability even within these variant classes. *BRCA2* c.9976A>T has a nonsense consequence but occurs close to the 3' end of the gene and is not regarded as significantly increasing the risk of breast or ovarian cancer.[76] An increase in the number of genotype-phenotype correlations such as this in cancer predisposition syndromes will be valuable for clinical management and might be expected as technological advances prompt greater numbers of individuals to undergo diagnostic testing.

1.7 - Impact of next generation sequencing on cancer predisposition gene variant identification in the clinic

NGS assays have had widespread implications for CPG variant identification in clinical settings. The most frequent group of assays applied by diagnostic services target (through PCR and/or selective pull-down) multiple genes potentially relevant to the patient's phenotype and are often referred to as

gene panels. Pathogenic variants in genes hitherto thought to be unrelated to the phenotype will not be detected through this method. The likelihood of this reduces as the number of tested genes increases and some panels aim to comprehensively cover all known CPGs. A yet more agnostic approach is that of exome or genome sequencing, where data relating to genes of interest can be selectively and flexibly analysed in a "virtual panel" technique and stored for future interrogation if new information regarding pertinent genomic regions becomes available.

The broadened scope of genetic analysis in clinical settings made possible by NGS technologies provides great opportunity to identify more individuals with previously unidentified cancer predisposing variants. Detection of variants in known CPGs in greater numbers of individuals allows more accurate characterisation of the phenotype associated with them in terms of tumour spectrum and penetrance. This begets the potential to reduce the aforementioned ascertainment biases associated with narrower access to testing, particularly when variants are found in patients with phenotypes previously considered uncharacteristic for aberrations at the locus in question.

1.8 - Clinical utility of cancer predisposition variant identification

Identification and characterisation of CPGs through research studies has produced opportunities to predict risk based on genetic factors elucidated by testing in clinical settings. Genetic testing may be diagnostic for individuals who have an existing cancer diagnosis and where an explanation is sought. Alternatively, predictive testing aims to assess risk in an unaffected individual through identification of relevant genetic variants. These are generally those that have previously been found in another family member but wider application of genetic analysis is likely to lead to more predictive testing where a variant has not been seen in a relative (e.g. in cases of adoption or deceased parents).

Even with the possibilities produced by NGS, resource constraints still limit the range of cancer patients that can be investigated. This is not only due to sequencing costs but also computational capacity, data storage, analytical time and sample availability. Prioritisation strategies are therefore often used to attempt to enrich for tumour predisposing variants (notwithstanding the associated ascertainment biases). Focus is often on a specific tumour type or clinical features suggestive of a specific syndrome but may also incorporate general indicators of cancer predisposition such as early age at diagnosis, occurrence of multiple primary tumours in the same individual and family history of neoplasia. Where family history is reported, the rationale for undertaking genetic testing may be stronger if a clustering of rarer tumours is observed as alternative causes are less likely. More ambiguity exists where common tumours cluster as this may be due to inherited predisposition or result from higher population incidence of the occurrent neoplasms, perhaps due to environmental factors. However, there is not a simple relationship between frequency of a specific cancer type and whether it is genetic or environmental in origin. An assessment of what proportion of cancer cases

were attributable to 14 preventable environmental exposures in the UK showed relatively low figures for many tumours with high population frequency including breast (26.8%) and colorectal cancers (54.4%).[77]

Whichever testing prioritisation strategy is chosen, successful elucidation of constitutional genetic factors that cause cancer predisposition can produce clinical utility in a number of ways.

1.8.1 - Information as therapy

Individuals undergoing genetic testing may value a diagnosis of a cancer predisposition syndrome independently of risk management or treatment as they may seek an explanation for frequently difficult personal and family histories of cancer. Negative results of diagnostic testing can provide reassurance although probands are often left with the possibility of unidentified pathogenic variants. A negative predictive test leads to greater confidence that the individual undergoing testing has a similar risk to the general population.

Much of the experience from genetic testing has been obtained via sequencing of *BRCA1* and *BRCA2* in clinical settings and a systematic review of psychological outcomes in women with a family history of breast cancer that underwent testing found a reduction in psychological distress for women receiving negative results and little change in those who received positive results.[78] A study of individuals undergoing predictive testing for *BRCA1* or *BRCA2* variants reported that 92% would recommend the process to others in the same situation.[79] One area of concern with predictive testing is the situation where some family members are found to carry a causative variant and others are not. An analysis of sibling dyads having predictive tests suggested some negative impact on relationship where the result was discordant between the two.[80] Any assessment of psychological benefits of genetic testing should be seen in the context of uptake, which, in the case of predictive testing, has been shown to be around half of eligible individuals for the conditions seen most commonly in the genetics clinic.[81,82] Those not pursuing testing may represent individuals who would not perceive as much benefit and future more widespread application of genetic testing could lead to more negative psychological sequelae in the absence of well-considered genetic counselling and consent processes.

Individuals consulting clinical services for assessment for a cancer predisposition syndrome may do so primarily to provide a genetic diagnosis in the family. This gives the opportunity for risk prediction and management in relatives even if prognosis is poor in the proband. An assessment of motivations for diagnostic testing in a series of colorectal cancer patients showed greater importance placed on information for relatives than desire to increase certainty regarding personal risk.[83]

Lastly, identification of pathogenic CPG variants in potential parents may facilitate reproductive decisions and produce the possibility to test for the variant in a foetus (prenatal diagnosis) or pre-implantation embryos (pre-implantation genetic diagnosis). These techniques are generally applied in severe (mainly non-neoplastic) childhood onset disorders and less frequently for cancer predisposition syndromes due to their frequently later onset and manifestations that are more amenable to risk mitigation strategies. However, a number of adult onset cancer syndromes are present on the list of conditions approved for pre-implantation genetic diagnosis by the Human Fertilisation and Embryo Authority[84] and a reportedly high proportion of individuals at risk of Lynch syndrome who regard prenatal diagnosis as ethically acceptable[85] suggest that uptake may increase in future.

1.8.2 - Clinical surveillance

A current mainstay of cancer predisposition syndrome management is regular clinical surveillance of at-risk tissue to identify tumours at an earlier and more treatable stage. A number of potential modalities exist for this purpose such as imaging, endoscopic examination and biochemical analysis, which are applied depending on the tissue or syndrome in question. Age range and frequency of surveillance investigations are guided by observational evidence from series of affected cases but the quality of this evidence can be compromised by rarity of a condition and/or ascertainment biases influencing which patients are included in studied cohorts.

Effectiveness of surveillance programmes is currently uncertain for most cancer predisposition syndromes but for conditions that have a higher incidence, larger screened cohorts can be assembled to provide greater clarity. A systematic review of Lynch syndrome screening, for example, showed reduction in colorectal cancer incidence and related mortality in screened (with regular colonoscopy) cases.[86] In rarer conditions, inference can be made from indirect information sources. Von Hippel Lindau disease leads to increased risk of a number of tumours including central nervous system haemangioblastoma, phaeochromocytoma and renal cell carcinoma. Protocols for surveillance are widely used but no prospective follow up data comparing screened with unscreened individuals exists. However, an increase in mean survival by 16.3 years has been observed among patients diagnosed after 1990, a time when systematic screening protocols were introduced.[87]

Surveillance programmes may be more straightforward where there are relatively few at-risk tissues to screen but many cancer predisposition syndromes lead to diverse tumour risks that can make execution of surveillance more complex and potentially less acceptable to patients. Li-Fraumeni syndrome is associated with a high risk of cancer that may arise from multiple organs and intensive, multi-modality screening regimens have been proposed in response to this.[88] Uncertainties surrounding effectiveness of these strategies has led, in many services, to a focus only on breast screening where greater confidence of utility exists. However, evidence is accumulating regarding the

benefits of whole body magnetic resonance imaging (MRI) and a 14% cancer detection rate from a one-off MRI has been reported in pathogenic *TP53* variant carriers.[89] A meta-analysis of this technique including 578 carriers reported a rate of 7% that was only contributed to by mostly non-breast cancers.[90] Promising figures such as these should be seen in the context of false positives and in the former study, 34% of 44 participants underwent further investigation for a lesion eventually diagnosed as benign with a corresponding figure of 24% in the meta-analysis.

Screening has more harmful potential (e.g. through unnecessary biopsy or surgery) where the penetrance of a CPG is low. In Hereditary Leiomyomatosis and Renal Cell Carcinoma caused by deleterious *FH* variants, only 15-20% of variant carriers develop kidney cancer but of those that do, many have reached an advanced stage with associated poor prognosis.[91] Difficult clinical situations such as this may be assisted by stratification of risk amongst CPG variant carriers, potentially based on the particular variant in the causative gene or other factors such as constitutional genetic modifers. Alternatively, acceptability, specificity and sensitivity of screening tests might be improved for low risk individuals by exploiting the phenomena of circulating tumour cells or DNA. Identification of specific markers of tumour development could facilitate potential future surveillance programmes based on blood sampling.

1.8.3 - Prophylactic surgery

In some syndromes where at-risk tissue is safely removable and non-essential, prophylactic surgery may be an effective preventative strategy. Influences on whether this is a reasonable option include extent of risk reduction, function (and loss thereafter) of the tissue in question and potential for complications following surgery. These factors need to be weighed against the efficacy and acceptability of surveillance strategies as an alternative. Prophylactic surgery can produce significant reduction in tumour risk and bilateral mastectomy in pathogenic *BRCA1* and *BRCA2* variant carriers is estimated to reduce the risk of breast cancer by around 90%.[92] Preventative oophorectomy has been reported to reduce ovarian cancer risk to a similar degree[93,94] but this procedure results in infertility and the requirement for hormone replacement in pre-menopausal women. Mastectomy may intuitively be regarded as having fewer negative consequences but negative psychological impact from this procedure can ensue.[95] In Hereditary Diffuse Gastric Cancer families, total prophylactic gastrectomy in pathogenic *CDH1* variant carriers is recommended but is associated with significant post-surgical morbidity from gastrointestinal symptoms.[96] The risk reduction provided by this procedure can be assumed to be significant but is difficult to quantify given the rarity of Hereditary Diffuse Gastric Cancer and the lower potential to assemble a series of controls (i.e. no surgery performed) with which to compare cancer incidence. A similar scenario exists for Familial Adenomatous Polyposis, where colorectal cancer risk[97] has been estimated to be at a level sufficient to warrant colectomy in all diagnosed cases, leaving few cases with an intact colon for further observation.

1.8.4 - Pharmacological management

Pharmacological prevention or treatment based on constitutional genetic status is in its infancy but it is hoped that this area will develop as molecular characterisation of syndromes and tumours accelerates.

Chemo-preventative strategies are seldom used in cancer predisposition syndromes but are an attractive proposition because side effects or economic cost are more likely to be outweighed by the high tumour risks involved. One of the more notable advances in this area has been the re-purposing of an established drug (aspirin) rather than development of a new agent. The observation of lower colorectal cancer rates in individuals taking aspirin prompted a trial in individuals with Lynch syndrome.[98] Here, daily aspirin was associated with an approximate 60% reduction in colorectal cancer incidence[99] and later guidelines indicated that aspirin use should be discussed with individuals from Lynch syndrome families.[100]

Pharmacological interventions in cancer predisposition syndromes may also be based on targeting a specific cellular aberration due to the causative constitutional variant. This area has received increasing attention in recent years but examples of current use remain infrequent. Vismodegib is an inhibitor of the hedgehog signalling pathway that is abnormally upregulated in basal cell carcinomas resulting from constitutional *PTCH1* variants and a second hit of the wild type allele (Gorlin syndrome).[101–103] The agent has been demonstrated to reduce basal cell carcinoma occurrence in Gorlin syndrome[104] but cost has prevented approval for use in the UK in either the hereditary or sporadic context.[105] A more widely used group of agents are poly ADP ribose polymerase (PARP) inhibitors for *BRCA1/2* related tumours, which are generally deficient in double stranded DNA repair by homologous recombination due to a second hit of the wild type allele. PARP inhibitors disrupt a different DNA repair mechanism (base excision repair), thus rendering tumour cells non-viable whilst sparing other cells where a second hit has not occurred and homologous recombination persists.[106]

1.9 - Multiple Primary Tumours

Multiple primary tumours (MPT) describes the scenario where two or more histologically distinct tumours that are not due to metastasis, recurrence or local spread are diagnosed in the same individual. These may be synchronous (diagnosed at the same time point) or metachronous (diagnosed months to years apart).

1.9.1 - Multiple primary tumours in the general population

The first description of MPT is attributed to Theodor Billroth in the nineteenth century.[107] It has been considered a rare phenomenon but has been observed more often as cancer survivorship has

lengthened.[108] Registry based studies have highlighted MPT as an increasingly frequent problem[109] with a key study observing 253,536 individuals diagnosed with cancer in Connecticut between 1935 and 1982 and reporting second primary neoplasms in 6.6% of them.[110] A more recent review of European cancer registries revealed that 6.3% of registered tumours were subsequent primaries[111] and 16% of incident cancers reported to National Cancer Institute (USA) in 2003 were diagnosed in patients with a previous cancer.[112]

MPT due to processes that are non-random can be indicated by a higher than expected incidence of second primaries in individuals previously diagnosed with cancer. In the Connecticut study, individuals with cancer had 1.3 times the risk of developing a cancer than individuals without a malignant diagnosis.[110] Relative incidence can be expressed as a standardised incidence ratio (SIR), which is a ratio of observed incidence and expected incidence in a corresponding population adjusted for risk factors such as age, sex and socioeconomic status. Population based studies have shown raised SIRs for a variety of concordant and discordant tumour types following a first primary and in a registry containing 633,964 cancer incidences, the SIR for any cancer was 1.3 in men with a previous malignancy and 1.6 in women. Some SIRs were below 1, suggesting lower incidence of cancer in individuals with certain malignant diagnoses. One explanation for this is that therapy for an initial primary may serendipitously treat a nascent cancer, particularly concordant tumours but potentially also discordant. Alternatively, poor prognosis associated with particular tumours may lead to less extensive surveillance and reduced probability of diagnosis of subsequent cancers before death occurs. For example the SIR for gastric cancer in men is 0.6 after 10-38 years.[113]

1.9.2 - Aetiology of multiple primary tumours

Multiple factors may contribute to the occurrence of MPT whose relative importance may be challenging to assess.

Correlation of number of stem cell divisions and cancer occurrence in different tissues has been interpreted as showing that variation in cancer incidence between tissues, and therefore many tumours, can be largely explained by mutagenic events that are not due to exogenous exposures or inherited factors.[114] The lifetime risk of cancer (excluding non-melanoma skin cancer) in the UK is estimated at around 1 in 2 for those born in the year 1960[115] and under this rationale, many of the tumours contributing to that figure would have little exogenous or constitutional genetic contribution. These might occur in the same individual if survival following a first diagnosis is of sufficient duration.

Follow up for cancer diagnoses can lead to the detection of second primaries that would not otherwise have been detected in the patient's lifetime, referred to as lead time bias. This situation does not

explain the aetiology of the neoplasms but influences the rate and spectrum of multiple primaries observed in a population. Second cancers may be identified due to systematic examination or imaging of tissue at risk of recurrence, for example through skin examinations after diagnosis of cutaneous malignant melanoma.[116] Surveillance imaging modalities might also include other organs in which cancers may be detected incidentally, as has been reported during follow up for pancreatic and prostate cancer with positron emission tomography/computed tomography.[117–119] Surgical intervention for a first primary may reveal a synchronous tumour that may have remained undiagnosed if an alternative management strategy had been chosen. Endometrial cancer can be diagnosed after total abdominal hysterectomy and bilateral salpingo-oophrectomy for ovarian cancer,[120] though it has been debated whether this particular pairing represents truly distinct primaries.

Radiotherapy or chemotherapy used to treat a first cancer may beget subsequent primary tumours. This can include non-cytotoxic drugs such as tamoxifen, which increases the risk of endometrial cancer following treatment for breast cancer.[121] Second cancers caused by treatment frequently occur many years after initial carcinogenic treatment occurred. Robust causative associations between therapies and subsequent neoplasms are difficult to delineate for a number of reasons. Poor survival from some initial tumour types means that subsequent primaries are less likely to be observed in individuals with that diagnosis because death may occur before they are reported. In addition, best practice treatment regimens often change over time and between centres. Collation of individuals with a particular diagnosis who are treated in the same manner may be challenging, especially if the tumour type in question is uncommon. Some treatment modalities for particular cancers may have only recently been adopted and carcinogenic effects might not have been observed yet. For example, renal cell carcinoma has previously been considered to be resistant to radiotherapy but more recent evidence suggests utility for this approach,[122] potentially increasing rates of radiation-related malignancies in renal cell carcinoma patients.

Despite these difficulties, a range of associations with treatment have been demonstrated. Histological or molecular examination of neoplasms may not reveal distinguishing features between treatment related and sporadic tumours in all cases but is useful in some scenarios. For example, leukaemias exhibiting microsatellite instability are more frequent where a tumour is therapy related but rare where leukaemia is diagnosed in the absence of a personal history of cancer.[123]

Patterns of treatment related cancer show differences dependent on whether chemotherapy or radiotherapy is used. Radiation related cancers generally occur ten years or more following exposure[124] and associations have often been reported by studies observing survivors of events such as the atomic bomb attacks in Japan in 1945[125] and Chernobyl nuclear accident in 1986.[126] Solid tumours such as those of the thyroid, lung, stomach, skin and connective tissue (sarcoma) are the most

frequently associated with radiation exposure[127] with sites reflecting tissue sensitivity and area of exposure. Haematological tumours such as leukaemias also occur at increased rates and may occur sooner after exposure.[128] The association of radiotherapy for Hodgkin's lymphoma and breast cancer is well established and has led to alteration in Hodgkin's lymphoma management with the intention of reducing radiation dosage to breast tissue.[129–131]

Malignancies due to chemotherapeutic agents are more frequently haematological and may occur following a relatively short post exposure time period. Alkylating agents (e.g. etoposide) can cause acute myeloid leukaemia that usually manifests after five to seven years. Leukaemias due to epipodophyllotoxins often have a three year latency period.[132] Chemotherapy can also lead to solid tumours, one example being dose responsive increased bladder cancer incidence after cyclophosphamide administration.[133]

Carcinogenic effects of treatment can be modified by a range of variables, perhaps most intuitively by dosage as higher levels of radiation or cytotoxic agents can produce greater potential for mutational events. Higher dosages might also lead to lower risk due to enhanced induction of cell death in clones with malignant potential.[123] Age at treatment may also be a modifying factor. If this is younger, there is likely to be a longer length of time in which further tumours can occur and a number of the known therapy-tumour associations have been found through follow up of children with diagnoses such as neuroblastoma.[134] Rather than simply more time to observe subsequent primaries, there is also evidence that second primary incidence at a given time point in follow up is lower in individuals where treatment exposure occurred at a later age.[124] Theoretical explanations include greater cellular proliferation at earlier ages that enhances clonal expansion of cells that have undergone tumourigenic genetic changes and increases the probability that further tumourigenic mutations will occur in daughter cells. Whilst systemic chemotherapy may affect a large variety of tissues accessible via the circulation, the pattern of carcinogenesis due to radiotherapy is modified by the field of treatment. Increased incidence of lung and oesophageal cancer, for example, are observed after radiotherapy for breast cancer.[135] Combination of therapeutic modalities may produce modifying effects. Doxorubicin used in Wilms tumour patients increases the risk of breast cancer following radiotherapy[136] and higher frequency of gastrointestinal malignancies has been reported following combined chemotherapy and radiotherapy for Hodgkin's lymphoma than would be expected if the risks from each modality were summed.[137] Constitutional genetic factors may also influence probability of subsequent tumours after treatment. Cancer predisposition syndromes can increase sensitivity to chemotherapy or radiotherapy as is seen for basal cell carcinomas after radiotherapy for medulloblastoma in Gorlin syndrome[138] and various radiation induced neoplasms in Li-Fraumeni syndrome.[139] Indirect modifying effects due to genetic factors are exhibited by cytochrome p450 enzyme gene variants, which increase or decrease blood levels of chemotherapeutic drugs through their effect on metabolism of some agents.[140]

Particular environmental exposures may increase the risk of more than one cancer type and can consequently account for some MPT cases. Smoking, for example, has a role in the aetiology or both aerodigestive tract cancers and lung adenocarcinoma and incidence of the former is increased following diagnosis of the latter.[141] Distinct environmental exposures may also contribute to MPT and some may be common enough to give rise to many individuals who experience multiple exposures. Smoking prevalence in adults is estimated at 20% in England[142] while obesity affects an estimated ~25%.[143] Multiplication of probabilities would indicate that ~5% of adults have both exposures but this assumes random distribution in the population, which is not necessarily true (e.g. smoking and alcohol consumption, both carcinogenic factors, are associated with one another[144]).

A role for constitutional genetic factors in the causation of MPT is indicated by increased incidence of second cancers in those with a family history of a corresponding neoplasm as it can be inferred that the increase is likely due to a shared heritable component. Studies arising from the Swedish Family Cancer Database have reported increased incidence of concordant and discordant second primaries in breast cancer cases with an affected parent compared with those without a parent diagnosed with breast cancer. As an example, the SIR (based on expected population incidence) for ovarian cancer following breast cancer was 2.0 in those with a family history of breast cancer and 1.7 in those without. The SIRs for acute lymphoid leukaemia were 12.7 vs 1.9 and 4.6 vs 3.0 for breast cancer.[145] Similarly, greater incidence of a second colorectal cancer has been observed among patients who have a first degree relative with that tumour type with a two-fold risk observed compared to non-familial cases.[146] Such observations suggest that inherited genetic factors contribute to the burden of second cancers in the general population, a proportion of which are monogenic.

Cancer predisposition syndromes form the focus of this thesis and can be suggested by clinical observations such as diagnosis of neoplasia at a young age or a family history of tumours (but not in cases due to *de novo* variants), particularly if histological concordance is present or if neoplasms are associated with a particular syndrome (e.g. colorectal and endometrial cancers in Lynch syndrome). Multiple tumours *per se* are also frequently taken as a clinical indicator and many predisposition syndromes are associated with a high frequency of the phenomenon. A number of syndromes that affect cutaneous areas are very frequently associated with multiple primaries. Xeroderma Pigmentosum causes multiple squamous cell carcinomas, basal cell carcinomas and melanomas in sun exposed areas.[147] Gorlin syndrome due to *PTCH1* variants also predisposes to basal cell cancers.[148] Neurofibromatosis type 1 and type 2 lead to, amongst other manifestations, multiple cutaneous neurofibromas and bilateral vestibular schwannomas respectively.[149,150] In practice, diagnosis and treatment of each tumour as a separate entity is more likely to occur in syndromes causing internal malignancies. Multiple cancers have been observed in 55% of 91 Li-Fraumeni cases with pathogenic

variants in *TP53*[151] and in 3% of Peutz-Jeghers syndrome cases with pathogenic variants in *STK11*.[152] Retinoblastoma is the predominant feature of the syndrome that carries its name but the full tumour spectrum includes extra-ocular cancers such as osteosarcoma, soft tissue sarcoma and melanoma. Observation of 1,852 bilateral retinoblastoma cases alive at one year following diagnosis showed a cumulative incidence of second primaries at 50 years of 47% and 38% with and without family history (of retinoblastoma) respectively.[153] Subsequent primaries are also a significant feature of cancer predisposition syndromes more commonly seen in clinical genetics departments. A study of 491 breast cancer cases carrying pathogenic *BRCA1* or *BRCA2* variants demonstrated ovarian cancer incidence of 12.7% for *BRCA1* and 6.8% for *BRCA2*.[154] In an analysis of 127 endometrial cancer patients with pathogenic variants in mismatch repair genes associated with Lynch syndrome, 48% had developed colorectal cancer at 20 years following initial diagnosis.[155] Given associations such as these, many patients with MPT will be referred for clinical evaluation with the intention of elucidating the causative CPG variant using genetic testing.

Chapter 2 – Methods applicable to multiple sections

The methods outlined in this chapter are applicable to investigation discussed in multiple chapters in this thesis. Methods specific to particular analyses are discussed in the relevant chapters.

2.1 - Study participants

Study participants were invited for recruitment through identification by clinical genetics services or by participation in previous research studies. The criteria for invitation were the development of two primary tumours by age 60 years or three primary tumours by age 70 years. Individuals with a single primary could also be included if they had a first degree relative who fulfilled these criteria. Most participants were eligible for recruitment on the basis of multiple malignant tumours but benign neoplasms could also be taken into account. A breakdown of phenotype and how eligibility criteria were fulfilled for each analysis can be found in the methods section of the chapter pertaining to that analysis. In each family, there was a clinical suspicion of a cancer predisposition syndrome but routine genetic assessment/testing had not identified a constitutional molecular genetic diagnosis to fully explain the tumour phenotype at the time of recruitment. Tumours in the same tissue type and organ were considered separate primaries if, in the case of paired organs, they occurred bilaterally or if the medical record clearly denoted them as distinct. International Agency for Research on Cancer guidance for defining separate primaries were also used.[156]

All participants gave written informed consent to participate in the National Institute of Health Research (NIHR) BioResource Rare Diseases (BRIDGE), Molecular Pathology of Human Genetic Disease (HumGenDis), and/or Investigating Hereditary Cancer Predisposition (IHCAP) studies. The NIHR BioResource projects were approved by Research Ethics Committees in the UK and appropriate ethics authorities in non-UK enrolment centres. Ethical approval for HumGenDis and IHCAP was given by South Birmingham and East of England, Cambridgeshire and Hertfordshire Research Ethics Committees respectively.

2.2 - Tumour labelling and classification

Initially, each tumour reported by recruiters or detected in the medical record was labelled with a topographical and morphological code based on the International Classification of Diseases for Oncology, Third Edition.[157] Selected codes were the most specific possible given the information available e.g. the morphological code chosen for breast cancer could have been “Infiltrating duct carcinoma” (8500/3) if a histology report was provided or “Neoplasm, malignant” (8000/3) if only the descriptor “breast cancer” was provided by the recruiting clinician.

In order to provide phenotypic groups for data analysis and results interpretation, tumours that occurred in participants were subsequently binned into categories on the basis of the initial coding. Tabulation of occurrent tumours pertaining to each analysis performed is referred to in the section

describing that analysis. Bins were generally named on the basis of topographical site. Tumours were assigned to such bins if they occurred at the specified sites unless there was evidence of a histological subtype that wouldn't be clinically described by the site-based term (e.g. medullary thyroid cancer would not be included in the "thyroid" bin but papillary thyroid cancer or "thyroid cancer" would be included). If a tumour type was not well described by a purely site-based label, a bin was created with a more specific term (e.g. paraganglioma, gastrointestinal stromal tumour and non-melanoma skin cancer). Haematological tumour bins were labelled according to cell lineage (e.g. lymphoid, myeloid).

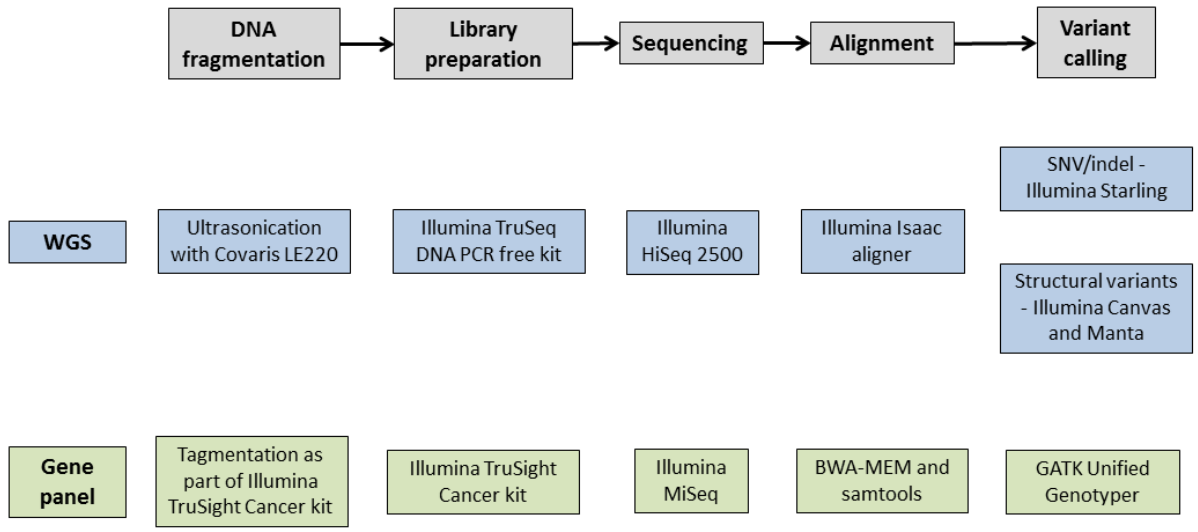
2.3 - DNA samples

DNA from lymphocytes was either obtained from DNA stored in diagnostic laboratories attached to clinical genetics centres or extracted from newly obtained blood samples. DNA extraction from blood was performed by the East Anglian Medical Genetics Laboratory using a Flex Star automated DNA extraction instrument (Autogen, Holliston, MA, USA). Some extractions from blood were performed by the Cambridge Translational Genomics Laboratory using a guanidine and precipitation-based methodology.

2.4 - Sequencing

Massively parallel sequencing was performed on blood DNA samples using whole genome sequencing (WGS) and a gene panel assay of cancer predisposition genes. The key steps in these processes are described in Figure 2.1.

Figure 2.1 - Key sequencing steps



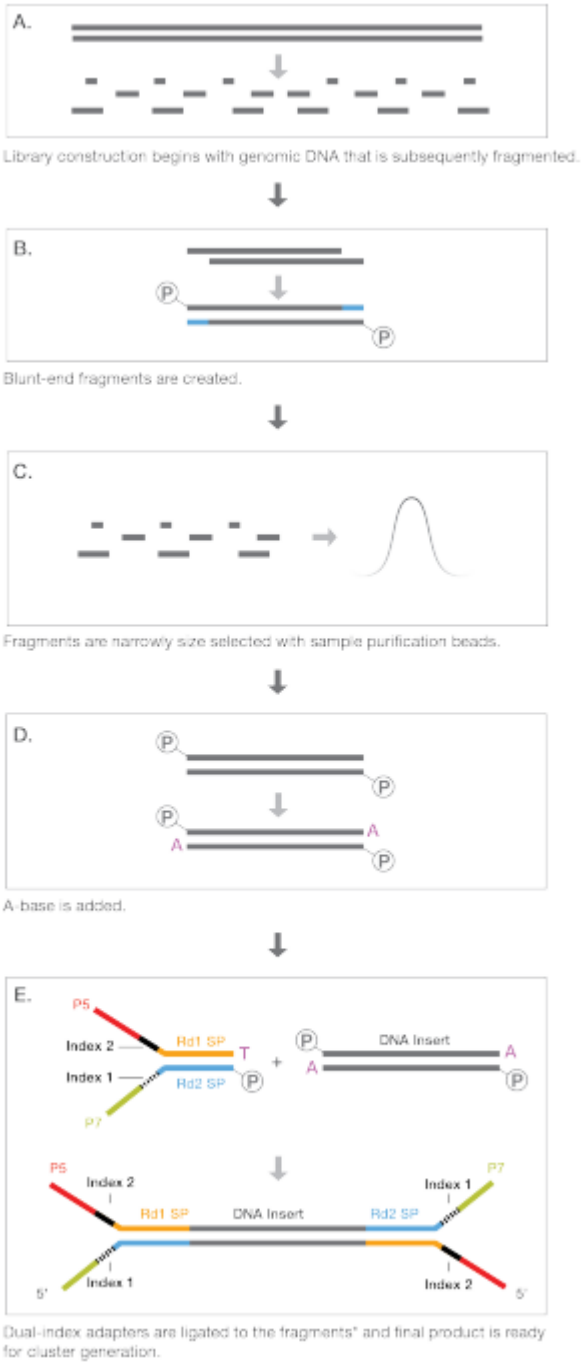
2.4.1 - Whole genome sequencing and bioinformatic processing of sequencing output

WGS and bioinformatic processing to produce variant call format (VCF) files was performed on samples from study participants as part of, and according to protocols devised by, the BRIDGE study.

DNA samples were checked for adequate concentration (30 ng/μl in 110 μl) with the PicoGreen assay (ThermoFisher, Waltham, MA, USA) and DNA degradation with gel electrophoresis. Purity was checked (adequate measurement optical density 260/280 1.75-2.04) with a Trinean DropQuant system (Trinean, Pleasanton, CA, USA). Samples passing quality control checks were shipped on dry ice to the sequencing provider (Illumina Inc., Great Chesterford, UK). Further quality controls were performed by the sequencing provider with a further check for adequate DNA concentration (30 ng/μl) and use of a microarray assay to ensure that samples were able to generate high quality genotyping results (Illumina Infinium Human Core Exome microarray). If samples were observed to have a repeated array genotyping call rate <0.99 or high levels of cross-contamination they did not go forward for WGS. The genotyping data were also used for sample identification before data delivery.

0.5μg of the DNA sample was fragmented using Covaris LE220 (Covaris Inc., Woburn, MA, USA) to obtain an average size of 450 base pair (bp) DNA fragments. DNA samples were processed using the Illumina TruSeq DNA PCR-Free Sample Preparation kit (Figure 2.2, Illumina Inc., San Diego, CA, USA) on the Hamilton Microlab Star (Hamilton Robotics, Inc., Reno, NV, USA). The final libraries were checked using the Roche LightCycler 480 II (Roche Diagnostics Corporation, Indianapolis, IN, USA) with KAPA Library Quantification Kit (Kapa Biosystems, Inc., Wilmington, MA, USA) for concentration.

Figure 2.2 - Illumina TruSeq DNA PCR-Free library preparation[158]



Libraries were sequenced with an Illumina HiSeq 2500 instrument with three different read lengths over the course of the project (February 2014 to June 2017). These were 100 bp (377 samples, three lanes used), 125 bp (3,154 samples, two lanes used). Some samples were sequenced with 150 bp reads (9,656 samples) on a single lane of an Illumina HiSeq X instrument. These numbers relate to all samples sequenced for the BRIDGE study and not just those that were in the multiple primary

tumours arm. The minimum coverage was 95% bases at 15X per lane and no more the 5% of insert sizes could be less than double the read length. The mean coverage achieved for 100, 125 and 150 bp read length was 41.4X, 37.9X and 35.3X respectively with a mean percentile of coverage of 31.0X, 25.7X and 26.2X. 90% of the utilised reference genome was covered at ≥19X in all samples.

Files containing sequencing data were delivered to and stored by the University of Cambridge High Performance Computing Service. FASTQ files were generated by HiSeq Analysis Software v2.0 (Illumina Inc., San Diego, CA, USA). Read alignment to GRCh37 was performed using Illumina Isaac aligner version SAAC00776.15.01.27.[32] Single nucleotide variants and indels were called from resulting binary compressed sequence alignment map (BAM) files using Illumina Starling software version 2.1.4.2. Output was in VCF and genome VCF format (gVCF), the latter of which contains information regarding coverage, alignment quality and other factors that contribute to a PASS filter at non-variant positions. gVCF files allow assessments of quality parameters at sites across samples to inform exclusion of problematic loci.

To identify sample duplication, a genotyping array was utilised to estimate kinship between samples. This assay incorporated a subset 8,872 single nucleotide polymorphisms (SNPs) randomly selected from those included on Roche microarrays for assessing kinship (Roche, Basel, Switzerland). Assessments of kinship using resulting data were performed using PLINK[159] and an output indicating a high degree of kinship prompted investigation as to the reason. Samples demonstrated to be duplicates or where the cause could not be determined led to the exclusion of one of the samples with inferior WGS data quality.

Measures were also taken to exclude samples on the basis of inadequate variant quality. Samples were removed if more than 5% of sites did not pass quality filters in the gVCF or if the ratio of observed transitions to transversions (which can be used to assess accuracy of single nucleotide variant calls[160]) fell outside of the interquartile range of values observed in the relevant sequencing batch. Additionally, exclusions were made if an inadequate proportion (<99.45%) of variant calls from common single nucleotide variant positions passed quality filters. Common variants were defined as those with a population specific minor allele frequency of >5% in gnomAD.[161] Contamination of samples by other DNA samples was also checked using verifyBamID software[162] and exclusion made if estimated contamination exceeded 3%. Sites associated with consistently poor quality calls across retained samples were excluded from all retained samples. Exclusion was based on overall pass rate that, for a given site, describes the proportion of samples where a call was possible multiplied by the proportion of those calls that passed quality filters. A threshold of overall pass rate of 0.99 was utilised.

Annotation of variants was performed according to the downstream analysis used in this project (see relevant chapters) but frequently utilised UK10K[163] allele frequency information that was added to variants by BRIDGE annotation pipelines.

Structural variant calling algorithms Canvas version 1.1.0.5[164] and Manta version 0.23.1[165] were also applied to the data. The former detects copy number variation based on sustained increases or decreases in sequencing read counts along genomic regions and is best suited for variants that exceed 10kb in length. The latter predicts inversions, translocations, tandem duplications, insertions and deletions based on the presence of split reads and/or evidence from paired reads and is designed to detect variants between 50bp and 10kb. Separate files containing calls corresponding to all structural variant modalities were provided for analysis.

Ethnicity and relatedness to other sequenced samples was estimated using a further SNP array-based strategy incorporating 292,878 variant sites used by the HumanCoreExome-12v1.1, HumanCoreExome-24v1.0 and HumanOmni2.5-8v1.1 genotyping arrays. This number was reduced to unlinked, high quality SNPs used for analysis following exclusions. SNPs were excluded if there was a missing genotype in at least one sequenced individual, if the minor allele frequency was <0.3 amongst sequenced individuals, if more than two alleles had been observed in sequenced individuals or in 1000 Genomes Phase 3 data (to assist with coding of genotypes),[166] if the overall pass rate (see above) for a site was <0.99, or if assessment with PLINK[159] indicated linkage disequilibrium between pairs of SNPs ($r^2 > 0.2$). 32,875 SNPs passing these filters were considered in a principal component analysis of unrelated individuals (defined using the KING R package[167]) in the 1000 Genomes Project Phase 3 performed using PC-AiR and PC-Relate functions of the GENESIS R package.[168] The resulting kinship matrix was analysed by PRIMUS software to produce a final set of unrelated individuals with pre-designated population of origin as part of 1000 Genomes annotation, forming the basis of partition into non-Finnish Europeans, Finns, Africans, South Asians and East Asians. Genotypes from individuals sequenced by the BRIDGE project were subsequently projected on to the 1000 Genomes principal components and the most likely ethnicity calculated on the basis of likelihood of the projected data assuming each of the five ethnicities. A numerical assessment of degree of familial relatedness was provided by a similar process which merged BRIDGE data with 1000 Genomes data (to produce greater genetic diversity for principal component analysis) and executed PC-Relate on input data.

2.4.2 - Gene panel sequencing and bioinformatic processing of sequencing output

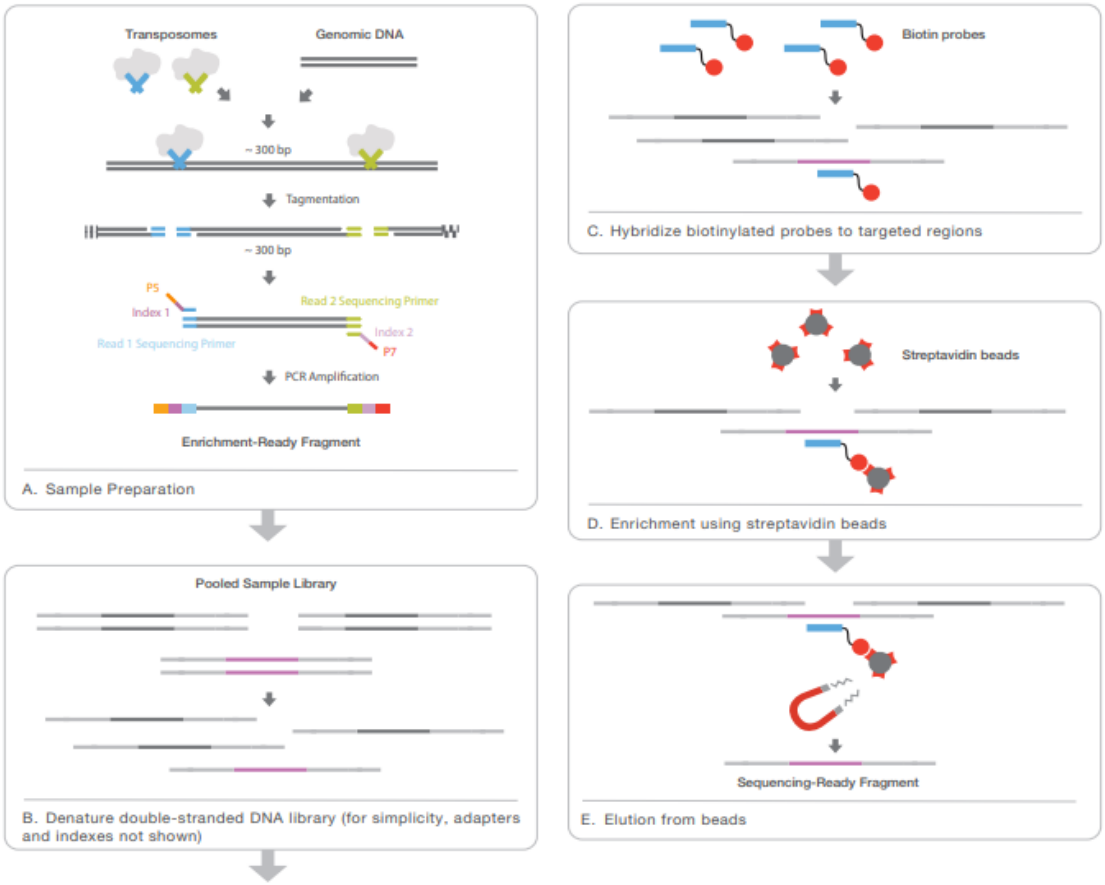
Gene panel-based sequencing and bioinformatic processing to produce VCF files was performed by colleagues in the Department of Medical Genetics, University of Cambridge.

The Illumina TruSight Cancer panel (Illumina Inc., San Diego, CA, USA) is the result of collaboration between the Institute of Cancer Research and Illumina to produce an assay that sequences a comprehensive collection of 94 cancer predisposition genes (Table 2.1). Library preparation from DNA samples was performed according to the manufacturer's protocol (Figure 2.3). Assessments of fragment size, quality and quantification were performed using a Bioanalyzer instrument (Agilent, Santa Clara, CA, USA).

Table 2.1 - Genes sequenced by Illumina TruSight Cancer panel

AIP	*CEBPA*	*FANCA*	*KIT*	*PRF1*	*SLX4*
ALK	*CEP57*	*FANCB*	*MAX*	*PRKAR1A*	*SMAD4*
APC	*CHEK2*	*FANCC*	*MEN1*	*PTCH1*	*SMARCB1*
ATM	*CYLD*	*FANCD2*	*MET*	*PTEN*	*STK11*
BAP1	*DDB2*	*FANCE*	*MLH1*	*RAD51C*	*SUFU*
BLM	*DICER1*	*FANCF*	*MSH2*	*RAD51D*	*TMEM127*
BMPR1A	*DIS3L2*	*FANCG*	*MSH6*	*RB1*	*TP53*
BRCA1	*EGFR*	*FANCI*	*MUTYH*	*RECQL4*	*TSC1*
BRCA2	*EPCAM*	*FANCL*	*NBN*	*RET*	*TSC2*
BRIP1	*ERCC2*	*FANCM*	*NF1*	*RHBDF2*	*VHL*
BUB1B	*ERCC3*	*FH*	*NF2*	*RUNX1*	*WRN*
CDC73	*ERCC4*	*FLCN*	*NSD1*	*SBDS*	*WT1*
CDH1	*ERCC5*	*GATA2*	*PALB2*	*SDHAF2*	*XPA*
CDK4	*EXT1*	*GPC3*	*PHOX2B*	*SDHB*	*XPC*
CDKN1C	*EXT2*	*HNF1A*	*PMS1*	*SDHC*	
CDKN2A	*EZH2*	*HRAS*	*PMS2*	*SDHD*	

Figure 2.3 - Illumina TruSight Cancer library preparation (taken from Illumina datasheet[169])



Libraries were sequenced with an Illumina MiSeq (Illumina Inc., San Diego, CA, USA). BCL files resulting from the sequencing were converted in FASTQ files using Illumina's bcl2fastq (Illumina Inc., San Diego, CA, USA). FASTQ files were checked for coverage and other quality control parameters using fastqc software. FASTQ files were aligned to the hg19 version of the reference genome using BWA-MEM[31] with default parameters and samtools[170] to produce a BAM file. Variants were called from BAM files using the Genome Analysis Tool Kit (GATK) Unified Genotyper algorithm.[33,171]

Chapter 3 – Multiple primary tumours in referral and registry-based series

Sections of this chapter discussing composition of a series of research participants with multiple primary tumours are based on corresponding sections of a previously published journal article (Whitworth et al).[172]

3.1 – Introduction

Research participants forming the basis of the studies presented in this thesis were individuals with multiple primary tumours (MPT) that were recruited via clinical genetics centres after referral for suspected cancer predisposition syndromes. Referrals to cancer genetics services are influenced by the relatively narrow range of cancer predisposition genes (CPGs) and well-defined syndromes that have historically prompted assessment. Indeed, previous analyses have recorded that referrals for breast or bowel cancer (associated with hereditary breast/ovarian cancer and Lynch syndrome) make up around 80% of the total.[173,174] However, inherited cancer syndromes as a whole can lead to a wide spectrum of tumours. Many affected individuals may not be assessed in the clinic but increasing sequencing capabilities of National Health Service genetics laboratories offers greater opportunity to do so. Although there are numerous epidemiological assessments of MPT in the literature,[109,111,175] reports often focus on risks following a specific initial cancer rather than a the relative occurrence of particular combinations.

To assess the nature of MPT combinations occurring in general populations, data from two cancer registries and a large treatment centre were obtained. Additionally, a series of MPT cases was ascertained through clinical genetics services that went on to be subject to sequencing analyses (herein referred to as the MPT series). This was compared with the registry series considered most representative of the population from which the MPT series was drawn to highlight differences that might influence the range of cancer predisposing genetic variants observed.

3.2 - Methods

Scripts used in these analyses are stored as an appendix in the form of a GitHub repository (https://github.com/jameswhitworth/Thesis-Elucidating_the_genetic_basis_of_multiple_primary_tumours-Scripts_appendix doi:10.5281/zenodo.1501206). They are denoted with the prefix "RA" (repository appendix) in the text and in the repository. Script RA3.1 was used for all collations, calculations and figures in this chapter.

3.2.1 - Collection and analysis of registry data

Data pertaining to individuals diagnosed with two or more cancers before the age of 60 years were obtained from three sources. The National Cancer Registration Service – Eastern Office (East Anglia (EA) Registry) covers a population in the UK of ~5.5 million[176] and provided data covering a period

from 2009-2014. Data were also obtained from the Netherlands Cancer Registry (Dutch Registry) covering a period from 1989-2014. Additionally, records with no time limit were interrogated from the Antoni Van Leeuwenhoek hospital (AVL), a major cancer treatment centre in Amsterdam, Netherlands. Data were filtered to only include information relating to tumours diagnosed at age 60 years or below.

Classification of tumours was based on International Classification of Diseases for Oncology (ICD-O-3)[157] topographical and morphological codes. Topographical codes were available for all tumours but some entries in the AVL data lacked a morphological code. In order to maximise the proportion of genuinely multiple primaries in the data, International Agency for Research on Cancer criteria[156] were applied. These criteria group sites and histological diagnoses that are considered to be equivalent in order to assist with classification. For a given individual, a maximum of one tumour from each topographical code grouping (the earliest to occur) was included unless any tumours at that same site were within a distinct morphological grouping. A final descriptive classification for each tumour was based on site and cell of origin as outlined in Table A1 (table predominantly describes tumours in MPT series but provides classification information for all tumours in registry/treatment centre data). Combinations of discordant cancers were then counted with individuals diagnosed with more than two tumours having multiple combinations assigned to them. For example, a history of tumours A, B and C would result in combinations A-B, A-C and B-C being recorded.

For tumours making up the collated combinations, possible indicators of a higher likelihood of a cancer susceptibility syndrome as a significant causative factor (rather than environmental exposures or chance) were noted. Although the most frequently diagnosed syndromes are associated with common tumour types, rare tumours may indicate a lesser role of chance as a predominant cause and it was noted whether the neoplasm was among the UK top five incident cancers (which make up 64% of all cancer diagnoses in the UK[177]). Heritability estimate was also noted for the occurrent tumours as a higher heritability estimate should increase the probability of genetic predisposition contributing to the tumours observed. Heritability describes the proportion of variance of a given phenotype that is attributable to inherited factors although it does not imply the relative role of numerous lower penetrance vs individual higher penetrance factors. For various cancer types, heritability has been estimated using statistical techniques that control or adjust for non-inherited factors such as environmental exposure, most notably through twin studies.[49,50] Estimates obtained from two such studies (Czene et al. 2002 and Mucci et al. 2016) were applied to tumours in this instance.

3.2.2 - Ascertainment and description of a multiple primary tumour series

A series of MPT cases was ascertained in order to study the molecular genetic basis of the tumours diagnosed in those individuals. 460 participants from 440 families were recruited through clinical

genetics services in the UK (442 cases), Greece (nine cases), Hong Kong (three cases), USA (three cases), Israel (two cases) and Ireland (one case). In each family there was a clinical suspicion of a cancer predisposition syndrome, but routine genetic assessment/testing had not identified a constitutional molecular genetic diagnosis at the time of recruitment. 435 individuals had developed MPT (defined here as ≥2 primaries by age 60 years or ≥3 by 70 years) while 25 had developed a single primary and had a first-degree relative with MPT. Tumour classification and counting of combinations was performed in the same manner as for the registry series.

3.2.3 - Comparison of Multiple Primary Tumour series with a population series

To consider how the tumour combinations in the MPT series differed from a general population, the combination frequencies were compared with the EA registry dataset as this was considered to be the most similar to the population from which the MPT series was drawn. Registry data recorded individuals with two cancers (or central nervous system (CNS) tumours) diagnoses before the age of 60 and only included tumours diagnosed before that age. Consequently, only combinations in MPT data of two malignant (or CNS) tumours occurring under age 60 were considered for this comparison (n=430). Two tailed Fishers exact tests were performed using the fish.test function in R version 3.4.3.[178]

3.3 - Results

3.3.1 - Registry and treatment centre series

The AVL, Dutch registry and EA series contained 4004, 1592 and 471 individuals respectively but information regarding sex was not included in the original data as obtained.

The most frequent individual tumour types are shown in Table 3.1 (also includes information for MPT series only including tumours diagnosed before age 60 described below). 8433 tumours were observed in the AVL series, in which breast cancer was the most common (19.2% total). Breast cancer was the second most frequent tumour in the Dutch registry (11.4% of 4,111 tumours) and EA series (17% of 989 tumours). The most frequent tumour in the Dutch registry series was cancer of the aerodigestive tract (14.3%) while the most frequent in the EA series was non-melanoma skin cancer (NMSC, 25.3%). Lung cancer did not make up ≥2% of the total in the EA series.

A large diversity of combination types existed in all the datasets (4,725, 3,274 and 560 respectively) with only a small number making up more than 2% of the total for each dataset (Table 3.2, also includes information for MPT series only including tumours diagnosed before age 60 described below). In the EA series, NMSC in combination with breast cancer (13.9% of total) and melanoma (11.4% of total) were twice as frequent as the third most frequent combination. Aerodigestive tract

cancer in association with lung cancer (6.1% of total) was most frequent in the Dutch registry series whilst breast cancer and melanoma made up the largest proportion of total combinations in the AVL series (5.1% of total). More frequent combinations are described graphically in Figures 3.1, 3.2 and 3.3.

Figure 3.1 - AVL series tumour combinations comprising >0.25% total (equivalent to >2 combinations in MPT series, see below)

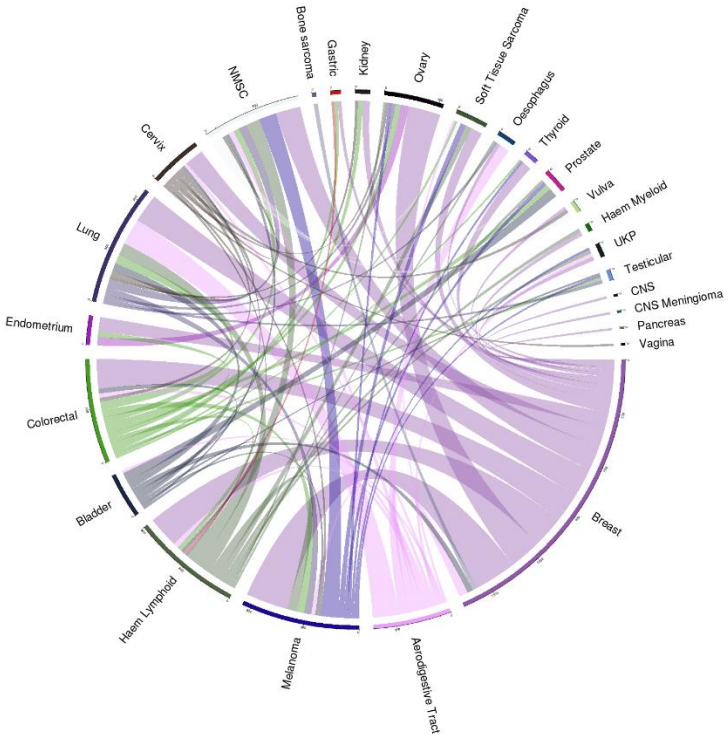


Figure 3.2 - Dutch registry series tumour combinations comprising >0.25% total (equivalent to >2 combinations in MPT series, see below)

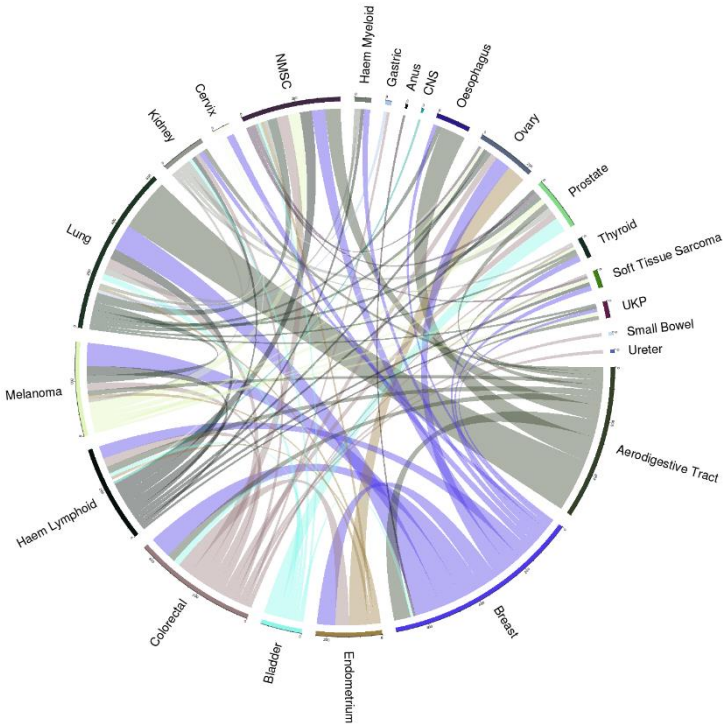


Figure 3.3 - EA Registry series tumour combinations comprising >0.25% total (equivalent to >2 combinations in MPT series, see below)

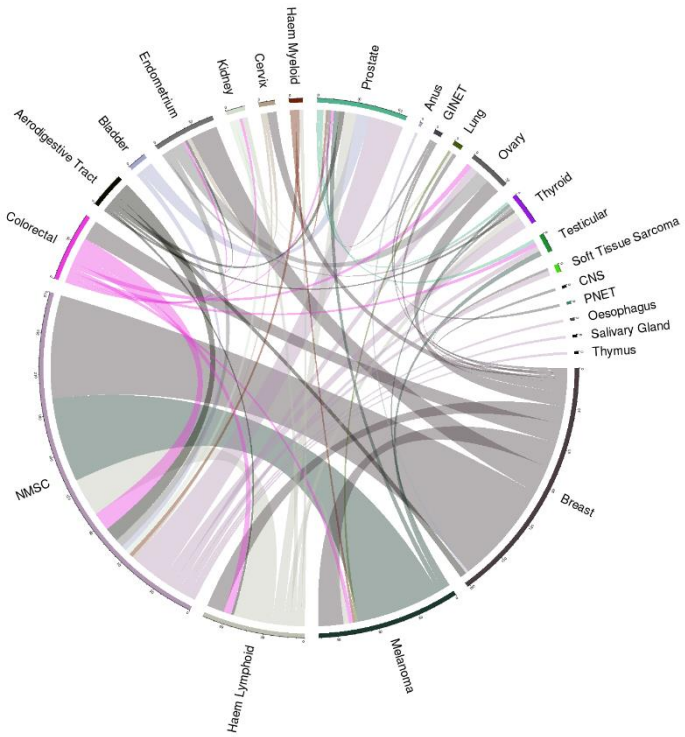


Table 3.1 – Most frequent tumour types in registry data and MPT (tumours under 60 only) series

Tumour type	Number
AVL	
Breast	1622 (19.2%)
Lung	699 (8.3%)
Colorectal	678 (8%)
Haematological lymphoid	647 (7.7%)
Melanoma	626 (7.4%)
NMSC	500 (5.9%)
Aerodigestive Tract	463 (5.5%)
Ovary	357 (4.2%)
Cervix	297 (3.5%)
Bladder	273 (3.2%)
Soft Tissue Sarcoma	238 (2.8%)
Endometrium	201 (2.4%)
Prostate	178 (2.1%)
Kidney	169 (2%)
Dutch registry	
Aerodigestive Tract	588 (14.3%)
Breast	467 (11.4%)
Lung	358 (8.7%)
NMSC	314 (7.6%)
Colorectal	310 (7.5%)
Haematological lymphoid	272 (6.6%)
Melanoma	242 (5.9%)
Endometrium	178 (4.3%)
Prostate	163 (4%)
Ovary	147 (3.6%)
Bladder	135 (3.3%)
Kidney	123 (3%)
Oesophagus	96 (2.3%)
East Anglia registry	
NMSC	250 (25.3%)
Breast	168 (17%)
Melanoma	93 (9.4%)
Haematological lymphoid	73 (7.4%)
Prostate	59 (6%)
Colorectal	52 (5.3%)
Endometrium	42 (4.2%)
Aerodigestive Tract	34 (3.4%)
Ovary	29 (2.9%)
Thyroid	27 (2.7%)
Bladder	17 (2%)
Testicular	17 (2%)
Kidney	16 (2%)
Lung	16 (2%)
MPT series (tumours under 60 only)	
Breast	221 (29.2%)
Colorectal	78 (10.3%)
Kidney	59 (7.8%)
Ovary	45 (5.9%)
NMSC	43 (5.7%)
Endometrium	40 (5.3%)
Thyroid	39 (5.1%)
Melanoma	38 (5%)
Haematological lymphoid	25 (3.3%)
Soft Tissue Sarcoma	13 (1.7%)
GIST	12 (1.6%)

Table 3.2 – Tumour combination types representing ≥2% total in registry data and MPT (only tumours diagnosed under 60) series

Combination	Number
AVL	
Breast-Melanoma	241 (5.1%)
Breast-Ovary	181 (3.8%)
Breast- Haematological lymphoid	179 (3.8%)
Breast-Colorectal	167 (3.5%)
Breast-Lung	163 (3.4%)
Aerodigestive Tract-Lung	149 (3.2%)
Breast-NMSC	142 (3%)
Breast-Cervix	108 (2.3%)
Melanoma-NMSC	100 (2.1%)
Dutch registry	
Aerodigestive Tract-Lung	201 (6.1%)
Breast-Lung	99 (3%)
Aerodigestive Tract-Oesophagus	94 (2.9%)
Breast-Melanoma	87 (2.7%)
Breast-Colorectal	83 (2.5%)
Endometrium-Ovary	70 (2.1%)
Breast-Endometrium	69 (2.1%)
East Anglia registry	
Breast-NMSC	78 (13.9%)
Melanoma-NMSC	64 (11.4%)
Haem Lymphoid-NMSC	29 (5.2%)
NMSC-Prostate	26 (4.6%)
Breast-Endometrium	21 (3.8%)
Breast-Melanoma	19 (3.4%)
Breast-Colorectal	15 (2.7%)
Colorectal-NMSC	14 (2.5%)
Breast- Haematological lymphoid	13 (2.3%)
MPT series (tumours under 60 only)	
Breast-Colorectal	29 (6.7%)
Breast-Ovary	23 (5.3%)
Breast-Endometrium	20 (4.7%)
Breast-NMSC	19 (4.4%)
Breast-Thyroid	19 (4.4%)
Breast- Haematological lymphoid	18 (4.2%)
Endometrium-Ovary	17 (4%)
Breast-Melanoma	14 (3.3%)

GIST – Gastrintestinal stromal tumour, NMSC – Non-melanoma skin cancer

3.3.2 - Multiple Primary Tumour series

460 individuals (106 (23%) males and 354 (77%) females) in 440 families had been diagnosed with 1,143 primary tumours distributed among 87 categories based on site and cell of origin (Table A1). The most frequent tumour types and combinations are shown in Table 3.3 and Figure 3.4. Breast cancer was the most frequent tumour representing 24.6% of the total with colorectal cancer (9.9%) the second. The most frequent combination type was breast and colorectal cancer representing 5.8% of the total combinations.

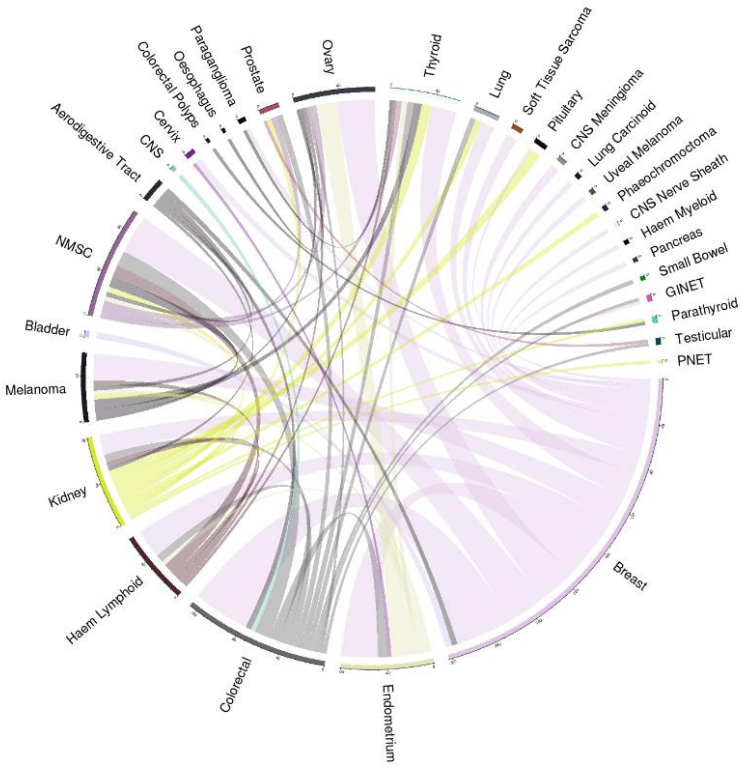
As per registry cases, the occurrence of any two discordant primaries in the same individual was considered as a tumour combination with a total of 883 combinations and 327 combination types observed (individuals with three or more discordant tumours would have multiple combinations). 206 (63%) combination types occurred once and 53 (16.2%) occurred twice. The 68 (20.8%) combination types occurring three or more times are illustrated in Figure 3.4.

Table 3.3 - Most frequent tumours and combinations in MPT series

Tumour category making up >5% total (total n=1,143)	**Number**
Breast	281 (24.6%)
Colorectal	113 (9.9%)
Kidney	83 (7.3%)
NMSC	67 (5.9%)
Ovary	58 (5.1%)
Tumour combination making up >1% total (total n=883)	**Number**
Breast-Colorectal	51 (5.8%)
Breast-NMSC	35 (4.0%)
Breast-Ovary	34 (3.9%)
Breast-Endometrium	33 (3.7%)
Breast-Haematological lymphoid	26 (2.9%)
Breast-Melanoma	24 (2.7%)
Breast-Thyroid	23 (2.6%)
Endometrium-Ovary	19 (2.2%)
Breast-Kidney	18 (2.0%)
Colorectal-NMSC	14 (1.6%)
Breast-Lung	12 (1.4%)
Haematological lymphoid-NMSC	11 (1.2%)
Breast-Soft Tissue Sarcoma	10 (1.1%)
Colorectal-Endometrium	9 (1.0%)
Kidney-Pituitary	9 (1.0%)
Kidney-Thyroid	9 (1.0%)
Melanoma-NMSC	9 (1.0%)

NMSC – Non-melanoma skin cancer

Figure 3.4 - MPT series tumour combinations occurring three or more times



Tumour combinations in all the series were assessed for characteristics suggestive of a greater likelihood of a significant inherited component (Table 3.4). Combinations where both tumours were not in the top five incident cancers and had a heritability estimate >20% made up 12.4% in the AVL series, 15.2% in the Dutch registry series and 4.8% in the EA series (Table 3.4). The equivalent figure in the MPT series was 11.4%, which reduced to 7.2% if only tumours under 60 were considered (see below)

Table 3.4– Tumour combination characteristics in registry data and Multiple Primary Tumour series

	MPT series	MPT series (only tumours under 60y)	AVL series	Dutch Registry	East Anglia Registry
Number of individuals	460	313	4004	1592	471
Number of discordant tumour combinations	883	430	4725	3274	560
≥1 tumour not among 5 most common	750 (84.9%)	366 (85.1%)	4067 (86.1%)	2864 (87.5%)	419 (74.8%)
2 tumours not among 5 most common	295 (33.4%)	120 (27.9%)	1321 (27.9%)	1033 (31.5%)	86 (15.3%)
One tumour with heritability estimate >20%	611 (69.2%)	274 (63.7%)	3532 (74.7%)	2675 (81.7%)	269 (48%)
Both tumours with heritability estimate >20%	174 (19.7%)	67 (15.6%)	1233 (26.1%)	1124 (34.3%)	50 (8.9%)
One tumour not among 5 most common and heritability estimate >20%	519 (58.8%)	229 (53.2%)	3030 (64.1%)	2333 (71.2%)	232 (41.4%)
Both tumours not among 5 most common and heritability estimate >20%	101 (11.4%)	31 (7.2%)	588 (12.4%)	499 (15.2%)	27 (4.8%)

3.3.3 - Comparison of MPT series (tumours under 60 only) with EA Registry series

To compare tumour combination distributions in the MPT series with a population-based dataset, the MPT series was subset to only include tumours diagnosed under the age of 60 years. This resulted in 313 MPT series individuals with 430 combinations (Figure 3.5), which were compared to 471 individuals with 560 combinations in the EA cancer registry data (Table 3.5). There was a significant difference (Fishers exact p value < 0.05) in tumour combination frequencies in 7/17 combination types that represented at least 1% of the MPT (tumours under 60 only) cohort total. Breast cancer in combination with ovarian, thyroid, lymphoid haematological, kidney cancer and meningioma were over-represented. Breast cancer in combination with non-melanoma skin was under-represented along with various other combinations involving skin cancers. Other less prominently over-represented tumour combinations were endometrium-ovary and kidney-thyroid.

Figure 3.5 - MPT series (tumours under 60 only) tumour combinations comprising >0.25% total (equivalent to >2 combinations in MPT series)

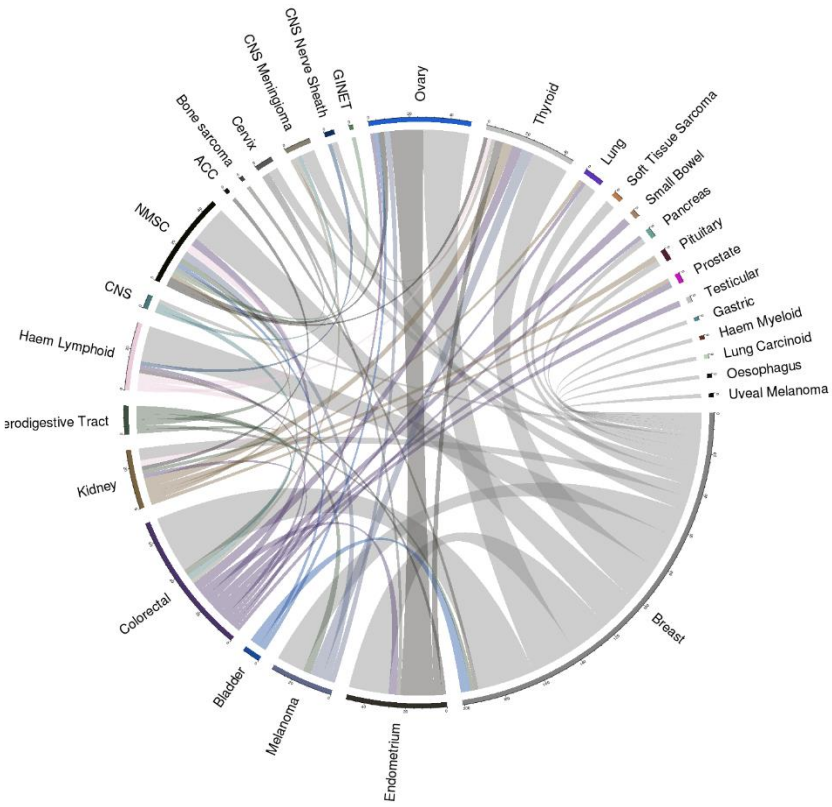


Table 3.5 - Comparison of MPT series (tumours under 60 only) with EA series

Combination	MPT count	MPT proportion (%)	EA count	EA proportion (%)	Difference in proportion MPT vs EA	Fishers exact p value (two tailed)
Breast-Colorectal	29	6.7	15	2.7	4	**0.00278**
Breast-Ovary	23	5.3	11	2	3.3	**0.00451**
Breast-Endometrium	20	4.7	21	3.8	0.9	0.52165
Breast-NMSC	19	4.4	78	13.9	-9.5	**<0.00001**
Breast-Thyroid	19	4.4	2	0.4	4	**0.00001**
Breast-Haem Lymphoid	18	4.2	13	2.3	1.9	0.10084
Endometrium-Ovary	17	4	10	1.8	2.2	**0.04809**
Breast-Melanoma	14	3.3	19	3.4	-0.1	1
Breast-CNS Meningioma	7	1.6	0	0	1.6	**0.00284**
Breast-Kidney	6	1.4	1	0.2	1.2	**0.04729**
Melanoma-Thyroid	6	1.4	2	0.4	1	0.08405
Breast-Lung	6	1.4	3	0.5	0.9	0.18776
Kidney-Thyroid	5	1.2	0	0	1.2	**0.01526**
Bladder-Breast	5	1.2	1	0.2	1	0.091
Colorectal-Thyroid	5	1.2	1	0.2	1	0.091
Breast-Soft Tissue Sarcoma	5	1.2	2	0.4	0.8	0.2498
Breast-Cervix	5	1.2	7	1.3	-0.1	1
Combinations not representing >1% total in MPT series (tumours Under 60) but comprising >1% total in EA series						
Melanoma-NMSC	4	0.9	64	11.4	-10.5	**<0.00001**
Haem Lymphoid-NMSC	4	0.9	29	5.2	-4.3	**0.00012**
NMSC-Prostate	0	0	26	4.6	-4.6	**<0.00001**
Colorectal-NMSC	4	0.9	14	2.5	-1.6	0.09153
Aerodigestive Tract-NMSC	3	0.7	11	2	-1.3	0.10941
Bladder-Prostate	0	0	10	1.8	-1.8	**0.00646**
NMSC-Thyroid	2	0.5	9	1.6	-1.1	0.12644
Haem Lymphoid-Prostate	0	0	9	1.6	-1.6	**0.00629**
NMSC-Ovary	3	0.7	6	1.1	-0.4	0.73911
Colorectal-Haem Lymphoid	2	0.5	6	1.1	-0.6	0.47736
Endometrium-NMSC	1	0.2	6	1.1	-0.9	0.14633

Haem, Haematological, NMSC – Non-melanoma skin cancer

3.4 - Discussion

3.4.1 - Registry and treatment centre-based data

To assess the nature of MPT at a population level, data was obtained from two cancer registries and a large cancer treatment centre.

The most frequent tumour types in those series broadly reflected established population frequency but notable differences were observed. NMSC accounted for over a quarter of tumours in the EA series but less than 8% in both the AVL series and Dutch registry. This may, as for other tumour types, reflect differences in reporting and recording by registries and in the case of the AVL series, pattern of referral to that centre. Lung cancer was infrequent in the EA series (2% total) but common in the

AVL and Dutch registry series. Lung cancer might be expected to be under-represented in multiple primaries cohorts as prognosis is poorer than for other common cancers where increased survival time increases the probability of further primaries. Possible explanations for the differences in lung cancer frequency between series include differences in lung cancer prognosis or detection/reporting of new primaries in terminally ill patients. Frequencies of all tumour types is likely to be influenced by the time period that the obtained data captured. Whereas the EA registry recorded 2009-2014, the Dutch registry went back to 1989. Changing incidence rates would therefore have influenced the cancer profile observed.

The vast majority of tumour combinations were comprised of combination types making up only a small proportion of the total. The more frequent tumour combinations broadly reflect those cancers that have a higher population incidence. Some recognised associations are also observed such as aerodigestive tract and lung cancer in the Dutch registry series, both associated with tobacco smoking.

A range of criteria proposed as suggestive of tumours being due to a cancer susceptibility syndrome were applied to the combinations and fulfilment of them recorded. Although the probability of such a syndrome conferred by these factors is not quantified, this suggested that combinations more likely to have a genetic aetiology exist in the population at appreciable rates. These figures were relatively consistent across the studied datasets. Whilst it is not known how many of these individuals were referred for clinical genetic assessment, this proportion may represent a group of individuals who would benefit from such assessment as testing capabilities develop.

3.4.2 - Comparison of Multiple Primary Tumour series with a population-based series

The MPT series was revised to only include tumours diagnosed under the age of 60 in order to make it comparable with the EA series. Striking differences were noted in frequencies of individual tumour types and combinations, likely reflecting common cancers with a significant hereditary component and for which genetic testing has been routinely available for a number of years. For example, breast cancer, while common in all series, made up close to a third of tumours in the MPT series. Kidney and colorectal cancers were also more frequent whilst NMSC, lung and aerodigestive tract cancers, which are generally not characteristic of cancer predisposition syndromes, were less frequent. Compared to EA registry cases, combinations such as breast-ovary (5.3% vs 2%) and breast-colorectal (6.7% vs 2.7%) are over-represented in the MPT (tumours under 60 only) series. Some of these cancers are sex specific, likely contributing to the uneven sex distribution in this series (although the sex breakdown of EA cases is not known). In some cases, specific tumour combinations may raise the possibility of a specific inherited cancer syndrome and prompt referral to genetics services (and hence the possibility of recruitment to the study). For example, the difference in breast-

thyroid frequency (4.4% in MPT series (tumours under 60 only) vs 0.4% in EA series) may be accounted for by suspicion of germline *PTEN* variants.

Chapter 4 – Analysis for variants in known cancer predisposition genes in a multiple primary tumour series

Sections of this chapter discussing interrogation of sequencing data from a series of research participants with multiple primary tumours for clinically relevant variants are based on a previously published journal article (Whitworth et al.).[172] The chapter is divided into three parts, the first of which (4.1) is concerned with detection of constitutional single nucleotide variants, indels and structural variants affecting cancer predisposition genes (CPGs). The second part (4.2) describes the formulation of a clinical scoring system to attempt to predict pathogenic variant carriers and the third part (4.3) discusses a search for mosaic CPG variants.

4.1 - Comprehensive analysis of known cancer predisposition genes in a multiple primary tumour series

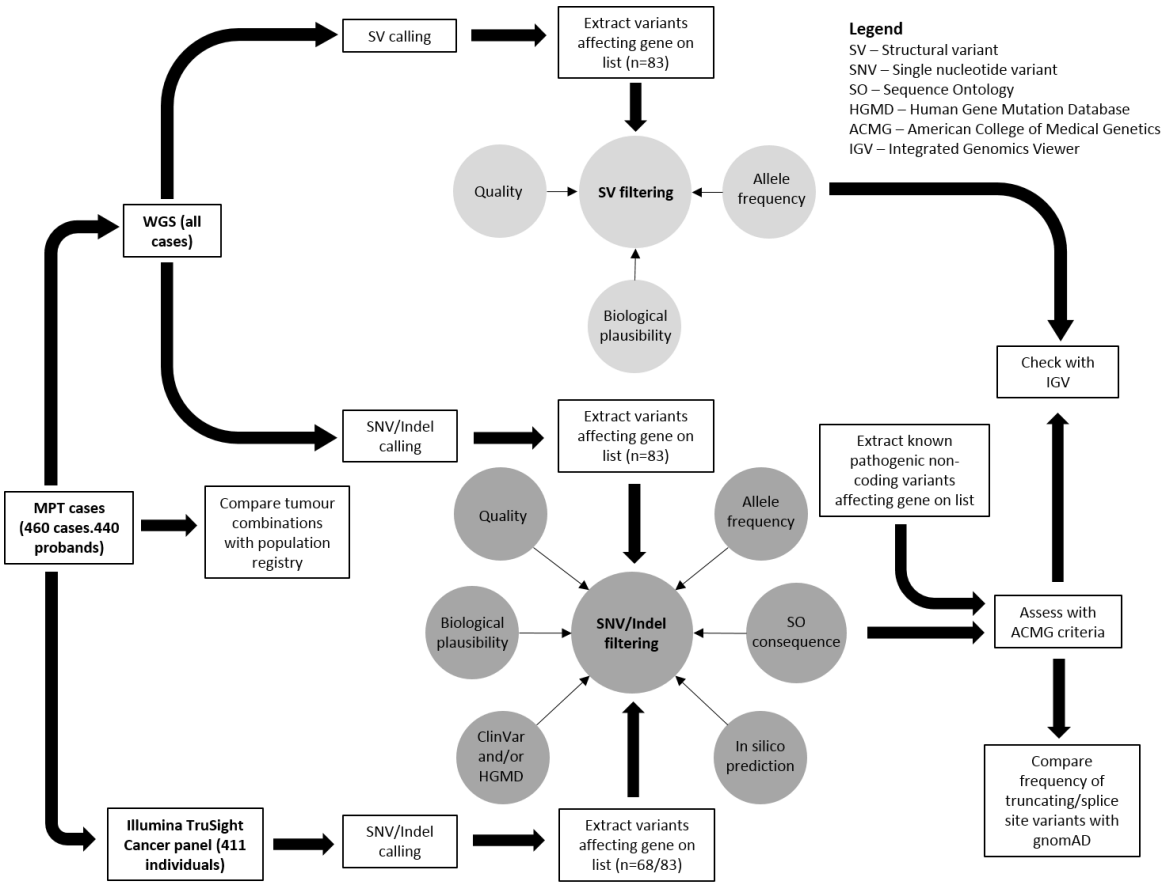
4.1.1 – Introduction

Clinical next generation sequencing (NGS) assays for possible inherited cancer predisposition generally target single genes or panels of CPGs but genome-wide analysis through whole exome sequencing (WES) or whole genome sequencing (WGS) is also possible. Though more expensive than WES, WGS should provide the most comprehensive analysis as it (a) can interrogate effectively all coding and non-coding areas of the genome, (b) provides more uniform read coverage compared to WES, particularly in areas where target enrichment/capture is difficult,[179,180] and (c) is able to detect a wide range of structural variations such as deletions, translocations, and inversions.[181] However, WGS is still in its infancy as a clinical diagnostic tool and few assessments of its application in hereditary cancer appear in the literature. Here, WGS has been applied to a large heterogeneous multiple primary tumour (MPT) cohort (n=460 incorporating 440 families) to investigate the potential role of comprehensive CPG analysis in this group.

4.1.2 - Methods

Workflow for the analysis is summarised in Figure 4.1. Scripts used in these analyses are stored as an appendix in the form of a GitHub repository (https://github.com/jameswhitworth/Thesis-Elucidating_the_genetic_basis_of_multiple_primary_tumours-Scripts_appendix doi:10.5281/zenodo.1501206). They are denoted with the prefix "RA" (repository appendix) in the text in and in the repository.

Figure 4.1 - Workflow for interrogation of whole genome sequencing data for clinically relevant variants



4.1.2.1 - Participants

460 participants from 440 families were recruited through clinical genetics services as described in Chapter 3. MPT was defined as ≥2 primaries by age 60 years or ≥3 by 70 years.

4.1.2.2 - Single nucleotide variant and indel identification in whole genome sequencing data and assessment (Script RA4.1)

Variants were extracted from variant call format (VCF) files if they were within a gene specified in a comprehensive list of 83 CPGs (gene list in Table 4.1). The gene list used for analysis was initially comprised of all genes listed in a 2014 review of CPGs[45] (n=114) and/or those sequenced by the TruSight Cancer panel (Illumina Inc., San Diego, CA, USA) (n=94, Table A2). Two additional more recently described CPGs were also included, namely *NTHL1* ([MIM:602656])[36] and *CDKN2B* ([MIM:600431]).[182] Genes were subsequently reviewed and filtered to produce a list that would be applicable to referrals to clinical cancer genetic services. Genes were included if deleterious variants affecting them are associated with adult onset tumours and if neoplastic lesions are likely to be a primary presenting feature. For example, *SOS1* was not included as although Noonan syndrome is

associated with increased neoplasia risk, other features of the condition are likely to prompt initial referral.

Table 4.1 - Gene list used for analysis (n=83)

AIP	*CDKN2A*	*EXT2*	*NF1*	*RAD51D*	*SMARCE1*
ALK[a]	*CDKN2B*	*FH*	*NF2*	*RB1*	*SRY*
APC	*CEBPA*	*FLCN*	*NTHL1*[b]	*RET*[a]	*STK11*
ATM	*CHEK2*	*GATA2*	*PALB2*	*RHBDF2*[a]	*SUFU*
AXIN2	*CYLD*	*HFE*[b]	*PDGFRA*[a]	*RUNX1*	*TGFBR1*
BAP1	*DDB2*	*HNF1A*	*PHOX2B*	*SDHA*	*TMEM127*
BMPR1A	*DICER1*	*KIT*[a]	*PMS2*	*SDHAF2*	*TP53*
BRCA1	*EGFR*[a]	*MAX*	*POLD1*	*SDHB*	*TSC1*
BRCA2	*EPCAM*	*MEN1*	*POLE*	*SDHC*	*TSC2*
BRIP1	*ERCC2*[b]	*MET*[a]	*POLH*[b]	*SDHD*	*VHL*
CDC73	*ERCC3*[b]	*MLH1*	*PRKAR1A*	*SERPINA1*[b]	*WT1*
CDH1	*ERCC4*[b]	*MSH2*	*PTCH1*	*SMAD4*	*XPA*[b]
CDK4[a]	*ERCC5*[b]	*MSH6*	*PTEN*	*SMARCA4*	*XPC*[b]
CDKN1B	*EXT1*	*MUTYH*[b]	*RAD51C*	*SMARCB1*	

a Genes considered as proto-oncogenes

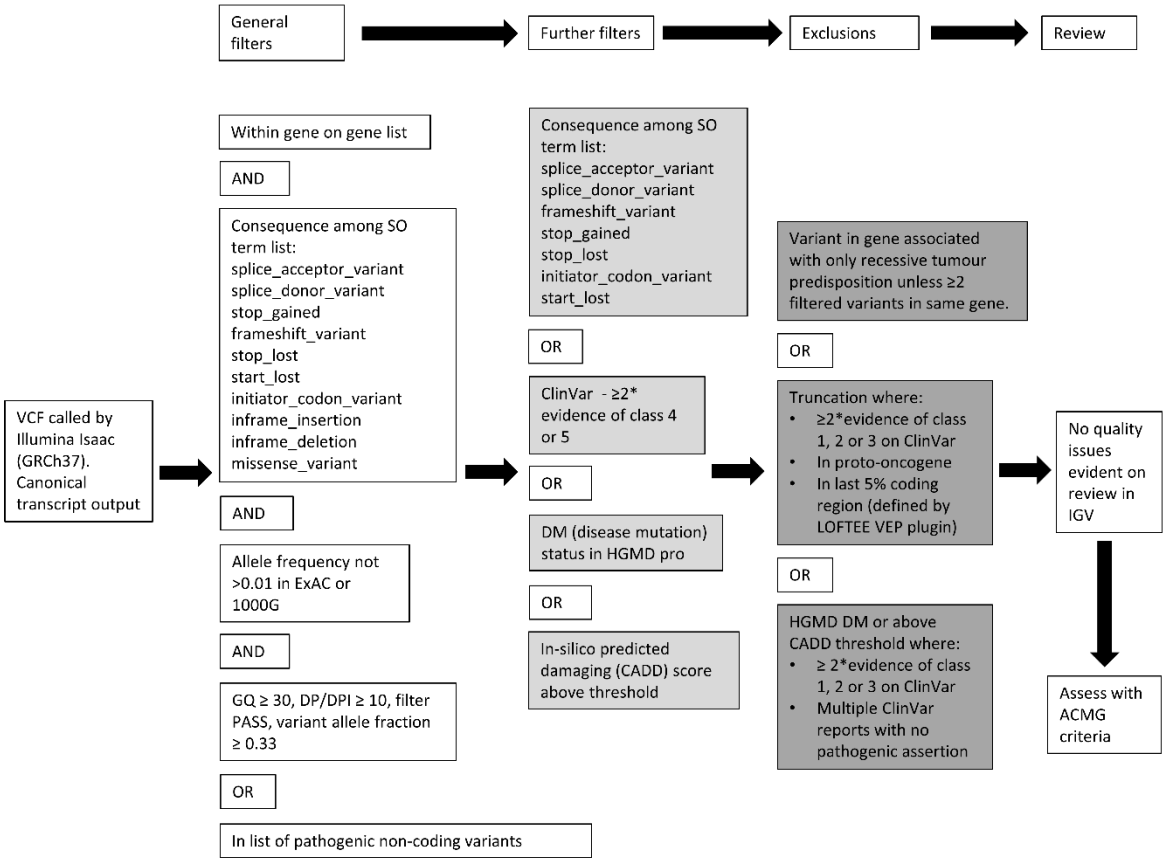
b Gene considered as associated with tumour predisposition in homozygous or compound heterozygous state only

For each gene on the gene list, the Ensembl canonical transcript identifier was selected by referencing gene-canonical transcript pairs provided by the Exome Aggregation Consortium (ExAC).[161] Canonical transcripts are defined according the following hierarchy: 1) longest Consensus Coding Sequence (CCDS)[183] translation with no stop codons, 2) Longest Ensembl/Havana merged translation with no stop codons, 3) longest translation with no stop codons and 4) if no translation, longest non-protein-coding transcript.[184] Lists of transcripts were then used to obtain GRCh37 coordinates for the protein coding regions within them with Biomart.[185] Coordinates were then used to produce BED files +/- 5 base pairs for use in filtering of VCF files. BED files were used in conjunction with bcftools (version 1.4) view[170] to extract variants in the corresponding regions and with FILTER PASS annotation (quality criteria as applied by the National Institute of Health Research BioResource Rare Disease (BRIDGE) project) from merged VCF files containing per chromosome variants called from BRIDGE WGS data (all sequenced individuals). Per chromosome files were merged with bcftools concat[170] and filtered with bcftools filter to remove variants if they failed to satisfy quality the quality criteria of GQ≥30 (phred scaled probability of the called genotype being incorrect), DP≥10 (number of reads covering the variant base/s 10 or greater) and variant allele fraction (VAF) ≥33%. The filtered merged VCF was then annotated with Variant Effect Predictor (VEP) version 90.[186]

In order to identify clinically relevant variants, resulting data were subject to a further range of filters (Figure 4.2) using the VEP filter script. Variants were excluded if they had an allele frequency above 0.01 in either ExAC[161] (all populations) or 1000 genomes project[166] (all populations). Variants were retained if the predicted consequence was among a list of sequence ontology (SO) terms indicating an effect on the protein product.

Filtered variants were considered for further review if the predicted consequence was among a list of SO terms indicating protein truncation and/or if there was evidence of pathogenicity in ClinVar[187] (≥2* evidence of pathogenic or likely pathogenic (P/LP) effect corresponding to multiple submissions with no conflicts as to assertion of clinical significance) or if the variant was assigned a disease mutation (DM) status in the Human Gene Mutation Database[188] (HGMD). In order to consider a subset of non-truncating variants that are predicted to be pathogenic by in-silico tools but do not appear in public databases, variants exceeding a phred scaled Combined Annotation Dependent Depletion (CADD)[189] score threshold of 34 were also retained for further review. CADD was selected for this purpose given that it incorporates a range of tools and consequently a number of lines of evidence. The threshold was chosen as the median of scores assigned to other variants (affecting any gene) deemed pathogenic according to the criteria described below. Identification of variants for retention due to CADD score alone was, therefore, done as a second variant filtering process after assessment of variants retained for other reasons.

Figure 4.2 - Filters applied to whole genome sequencing data – Single nucleotide variants and indels



ACMG – American College of Medical Genetics, HGMD- Human Gene Mutation Database, SO – Sequence Ontology, VCF – Variant call format

Sequence variants in non-coding regions such as introns that affected genes in the gene list would not be extracted from the original VCF files based on the strategy described as their SO consequence would not be within the utilised list. Therefore, a list of known pathogenic variants falling outside of exons or splice sites/regions was compiled using ClinVar and used to filter VCFs based on their genomic positions in a separate interrogation. Variants were incorporated in the list if they occurred in or near a gene on the list, were classified as near gene, non-coding RNA or untranslated region, and had ≥2* evidence of a pathogenic or likely pathogenic effect. This process produced only three known pathogenic variants to search for in the WGS data. Distant non-coding variants affecting gene function (e.g. enhancers) were not considered in the analysis described in this chapter.

Retained variants were subsequently excluded if their putative pathogenicity could be refuted by fulfilling one of the following criteria: 1) A predicted protein truncating variant where there was ≥2* evidence of a benign or uncertain effect in ClinVar, 2) A predicted protein truncating variant in a proto-oncogene in a list compiled based on literature review[45] (constitutional cancer predisposing

variants in proto-oncogenes are associated with gain of function variants so truncation of protein product is unlikely to increase tumour risk), 3) A predicted protein truncating variant affecting <5% of the canonical transcript (based on the LOFTEE VEP plugin), 4) A variant affecting a gene associated with only recessive tumour predisposition (as defined by literature review[36,45,190]) unless an individual appeared to harbour two filtered variants in the same gene, 5) An HGMD DM status variant or variant which exceeded the CADD score threshold where there was ≥2* ClinVar evidence of a benign or uncertain clinical effect or 1* evidence if there were multiple submissions without any containing a likely pathogenic or pathogenic assertion.

Variants passing filters were reviewed with Integrated Genomics Viewer[191] (IGV) to check for issues such as adjacent variants affecting the predicted consequence or variants being located at the end of sequencing reads. Pathogenicity was then assessed according to the American College of Medical Genetics (ACMG) criteria (Table 4.2),[192] which provides a framework to compile multiple weighted lines of evidence. Additionally, for each variant it was noted whether the corresponding individual had previously been diagnosed with a tumour typically associated with pathogenic variants in the relevant gene (according to Rahman,[45] the Familial Cancer Database,[190] or the original paper reporting the gene as a CPG). Validation of P/LP variants was carried out using data generated from Illumina TruSight Cancer panel (TCP) or by the BRIDGE project Sanger sequencing service according to standard protocols (if TCP data was unavailable).

Table 4.2 - American College of Medical Genetics criteria as applied to single nucleotide variant and indel analysis

Evidence of benign nature

Stand-alone evidence	**ACMG description**	**Application to present analysis**	**All variants tagged as not fulfilling criteria?**
BA1	Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium.	All variants fulfilling this criterion filtered prior to analysis.	Yes
Strong evidence	**ACMG description**	**Application to present analysis**	**All variants tagged as not fulfilling criteria?**
BS1	Allele frequency is greater than expected for disorder.	Uncertainties around prevalence and penetrance of inherited cancer syndromes prevent accurate assessment of this criterion. All variants are rare.	Yes
BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age.	Full penetrance at an early age not expected for inherited cancer syndromes caused by variation in genes considered.	Yes
BS3	Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing.	If variant present in HGMD with DM or DM? status, reviewed linked papers for functional studies. If variant annotated with PubMed ID by Variant Effect Predictor, reviewed listed articles. Loss of heterozygosity in tumour and/or evidence of RNA disruption considered.	No
BS4	Lack of segregation in affected members of a family.	Criterion not used due to lack of specificity of phenotypes and incomplete penetrance of inherited cancer syndromes considered.	Yes
Supporting evidence	**ACMG description**	**Application to present analysis**	**All variants tagged as not fulfilling criteria?**
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease.	Criterion fulfilled if no missense variants in the gene appear in HGMD (with DM status) or ClinVar with pathogenic assertion.	No
BP2	Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern.	Analysed variants not phased.	Yes
BP3	In-frame deletions/insertions in a repetitive region without a known function.	All filtered inframe deletions/insertions scored as PM4 following review. Therefore, none fulfil BP3.	Yes

BP4	Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.).	Fulfilled if CADD score (where given) 10 or below (corresponding to variant being outside top 10% predicted most deleterious variants).	No
BP5	Variant found in a case with an alternate molecular basis for disease.	Fulfilled for all variants due to alternative (non-genetic predisposition related) mechanism in all tumours.	No
BP6	Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation.	Fulfilled if any single report in ClinVar with benign/likely benign assertion.	No
BP7	A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved.	All variants fulfilling this criterion filtered prior to analysis.	Yes

Evidence of pathogenic nature

Very strong evidence	**ACMG description**	**Application to present analysis**	**All variants tagged as not fulfilling criteria?**
PVS1	Null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where LOF is a known mechanism of disease.	Fulfilled if variant had Sequence Ontology term (assigned by Variant Effect Predictor) consistent with one of these consequences unless within proto-oncogene.*	No
Strong evidence	**ACMG description**	**Application to present analysis**	**All variants tagged as not fulfilling criteria?**
PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change.	Fulfilled if missense variant leads to same amino acid change as a pathogenic missense variant as defined by ClinVar pathogenic or likely pathogenic with ≥2* evidence status.	No
PS2	De novo (both maternity and paternity confirmed) in a patient with the disease and no family history.	Incomplete penetrance may frequently lead to no family history in inherited cancer syndromes. Only one trio in this analysis (filtered variant was not de-novo).	Yes
PS3	Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product.	If variant present in HGMD with DM or DM? status, reviewed linked papers for functional studies. If variant annotated with PubMed ID by Variant Effect Predictor, reviewed listed articles. Loss of heterozygosity in tumour and/or evidence of RNA disruption considered.	No

PS4	The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls.	Number of variants and phenotypes in the series prevented use of this criterion.	Yes
Moderate evidence	**ACMG description**	**Application to present analysis**	**All variants tagged as not fulfilling criteria?**
PM1	Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation.	Fulfilled if occurred in Pfam[193] domain and relevant domain contains ≥1 pathogenic/likely pathogenic variants AND 0 benign/likely benign/VUS missense variants as defined by ClinVar ≥2* evidence status. Mutational hotspot criterion not used.	No
PM2	Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium.	Fulfilled if absent in either 1000 Genomes or ExAC based on Variant Effect Predictor annotation.	No
PM3	For recessive disorders, detected in trans with a pathogenic variant.	Analysed variants not phased. No compound heterozygotes for suspected recessive cancer predisposition identified among filtered variants.	No
PM4	Protein length changes as a result of in-frame deletions/insertions in a non-repeat region or stop-loss variants.	Fulfilled if variant has Sequence Ontology term predicted consequence and doesn't occur in repetitive region as defined by UCSC[194] repeat masker track.	No
PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before.	Fulfilled if missense variant is within the same codon as a pathogenic missense variant as defined by ClinVar pathogenic or likely pathogenic with ≥2* evidence status.	No
PM6	Assumed de novo, but without confirmation of paternity and maternity.	Unable to reliably assume de novo due to incomplete penetrance of inherited cancer syndromes considered.	Yes
Supporting evidence	**ACMG description**	**Application to present analysis**	**All variants tagged as not fulfilling criteria?**
PP1	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease.	Incomplete penetrance of considered inherited cancer syndromes and low number of participants per family prevented use of criterion.	Yes
PP2	Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease.	Fulfilled if variant occurs in gene with low rate of benign missense variation as defined by ExAC missense constraint metric <-3·09 (equivalent to observed vs expected p value 0·01) and ≥1 pathogenic/likely pathogenic missense variant in ClinVar with ≥2* evidence status.	No
PP3	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.).	Fulfilled if CADD score (where given) 30 or above (corresponding to variant being within top 0·1% predicted most deleterious variants).	No
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic aetiology.	Inherited cancer syndrome phenotypes considered not sufficiently specific for fulfilment.	Yes

PP5	Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation.	Fulfilled if any single report in ClinVar with pathogenic/likely pathogenic assertion or in HGMD with DM status.	No

ExAC – Exome Aggregation Consortium, HGMD – Human Gene Mutation Database, VUS – Variant of uncertain significance

4.1.2.3 - Single nucleotide variant and indel identification in gene panel data and assessment (Script RA4.2)

The variant filtering and assessment process described for WGS data was also applied to per individual VCF files containing variant calls made from TCP data.

4.1.2.4 - Structural variant identification and assessment (Script RA4.1)

Structural variant (SV) calls that were predicted to affect a gene on the gene list (n=83) were filtered and assessed according to the quality of the call, rarity of the variant, and biological plausibility of tumour predisposition caused by the variant (Figure 4.3).

Variant call files (txt format) provided by the BRIDGE project and containing calls for predicted deletions (separate files from Canvas and Manta), copy number gains (Canvas), translocations (Manta), duplications (Manta), inversions (Manta) and insertions (Manta) were used. Files were only available for 390 out of the 460 individuals included in the analysis of single nucleotide variants and indels. Variants were initially filtered by BRIDGE to retain those that were predicted to affect at least one exon, occurred at a frequency of less than 1% across all BRIDGE samples (n= 9110) in the data release utilised and were not associated with a flag introduced by Manta or Canvas indicating a low-quality call.

Genomic coordinates for genes of interest were based on gene start and gene end coordinates downloaded from Ensembl Biomart[184] (GRCh37 build). Manta annotation contains confidence intervals describing the range of bases surrounding the predicted SV coordinates that are likely to contain the true breakpoints of the variant. These values can be utilised to produce genomic positions corresponding to the minimum start, maximum start, minimum end and maximum end of any given SV. They were used in the identification Manta called SVs affecting regions of interest. SV calls were filtered using an R script according to the criteria outlined in Table 4.3 and minimum quality criteria of GQ ≥30 for Manta and QUAL ≥30 for Canvas.

Table 4.3 - Conditions used to identify structural variants

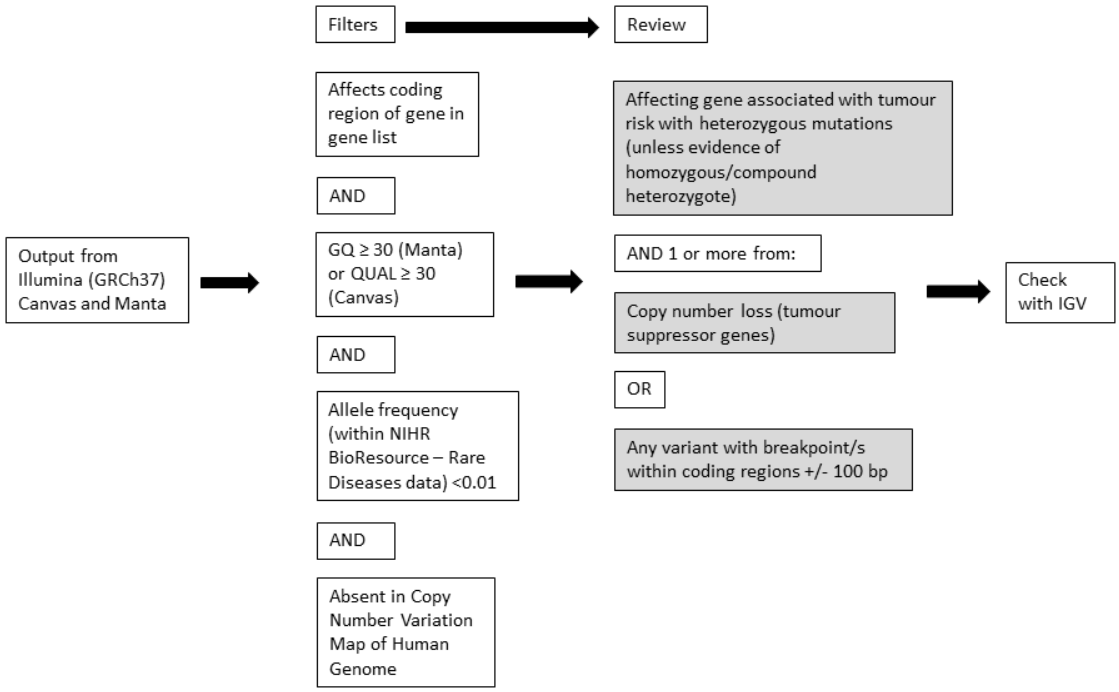
Structural variant modality	Conditions for structural variant call to fulfil
Deletion (Manta)	Max. start < gene start AND min. end > gene end OR Min. start > gene start AND max. end < gene end OR Max. start < gene start AND (min. end > gene start AND max. end < gene end) OR Min. end > gene end AND (max. start < gene end AND min. start > gene start)
Deletion (Canvas)	Start < gene start AND end > gene end OR Start > gene start AND end < gene end OR Start < gene start AND (end > gene start AND end < gene end) OR End > gene end AND (start < gene end AND start > gene start)
Copy number gain (Canvas)	Start < gene start AND end > gene end OR Start < gene start AND (end > gene start AND end < gene end) OR End > gene end AND (start < gene end AND start > gene start)
Translocation (Manta)	Min. start > gene start AND max. start < gene end OR Min. end > gene start AND max. end < gene end
Inversion (Manta)	Min. start > gene start AND max. start < gene end OR Min. end > gene start AND max. end < gene end
Insertion (Manta)	Min. start > gene start AND max. start < gene end OR Min. end > gene start AND max. end < gene end
Inversion (Manta)	Min. start > gene start AND max. start < gene end OR Min. end > gene start AND max. end < gene end
Duplication (Manta)	Max. start < gene start AND min. end > gene end OR Min. start > gene start AND max. end < gene end OR Max. start < gene start AND (min. end > gene start AND max. end < gene end) OR Min. end > gene end AND (max. start < gene end AND min. start > gene start)

Remaining variants were regarded as potentially pathogenic if they were predicted to affect a gene associated with tumour predisposition in the heterozygous state (unless there was evidence of homozygosity/compound heterozygosity) and fell into either of the following categories. 1) Copy number loss of coding regions of a tumour suppressor gene, 2) Predicted breakpoint disrupting a tumour suppressor gene. Copy number gain or breakpoints affecting proto-oncogenes was not taken as evidence of a clinically relevant SV given that known pathogenic variants in these genes tend to be a narrow range of missense variants exerting their effect through specific gain of function

mechanisms. It is difficult, therefore, to interpret increased gene dosage as equivalent to one of those variants.

Subsequently, these SV calls were reviewed with IGV and excluded if they occurred in the Copy Number Variation Map of Human Genome[195] (Hg19 stringent). Occurrence of tumours associated with disruption of particular genes in individuals harbouring suspected SVs was noted in the same manner as for single nucleotide variants and indels. BAM files corresponding to all suspected deleterious calls were reviewed in IGV. All SVs considered pertinent following filtering and assessment were confirmed with Sanger sequencing according to standard protocols. Inversions, translocations and tandem duplications were confirmed by sequencing across breakpoints while deletions could be confirmed by fragment size resulting from long range polymerase chain reaction if sequencing across the breakpoint was not possible. Validation was performed by colleagues in the Cambridge Translational Genomics Laboratory and, in one instance, the University of Cambridge Department of Medical Genetics.

Figure 4.3 - Filters applied to whole genome sequencing data – Structural variants



IGV – Integrated Genomics Viewer, NIHR – National Institute of Health Research

4.1.2.5 - Comparison of rate of truncating variants in Multiple Primary Tumour series vs gnomAD dataset (Script RA4.3)

To compare loss of function variant detection rates in the MPT cohort with a large scale WGS dataset unselected for neoplastic phenotypes, the gnomAD database[161] (downloaded February 2018) was

interrogated for variants occurring in the same set of 83 genes. Only truncating or splice site variants were considered for comparison purposes as these are less likely to be false positives and made up 52/63 (82.5%) (see results section) of the P/LP variants in the MPT cohort. Variants extracted from gnomAD were filtered and assessed as per those occurring in the MPT cohort. Given that the sex distribution of the MPT cohort was skewed towards females, frequency of variants assessed as P/LP was also calculated for males and females in both datasets. For the gnomAD data, the sex distribution (55.3% male, 44.6% female) was estimated by taking the sex specific mean allele count incorporating all positions in the gnomAD chromosome 1-22 VCF file and comparing the relative counts. In order to estimate gnomAD P/LP variant frequency as if sex distribution was equivalent to the MPT series (23% male, 77% female), a sex specific frequency based on the estimated sex distribution was applied to the estimated total number of gnomAD females (n=6929) and a reduced number of males (n=2064) that would achieve the desired proportion. The respective allele frequency estimates were then summed to provide a figure to compare with the MPT series.

4.1.2.6 - Calculation of sequencing coverage (Script RA4.4)

For BAM files from WGS and TCP data, coverage statistics for regions of interest were generated with samtools depth.[170] A BED file compiled using Ensembl BioMart[185] to represent translated exonic regions and splice sites of genes in the gene list was utilised.

4.1.2.7 - Statistical analysis

All statistical tests were performed using R version 3.4.3.[178] Pearson's chi-squared tests and students t tests were performed using the chisq.test and t.test functions respectively.

4.1.3 - Results

4.1.3.1 - Clinical characteristics and multiple primary tumour combinations

The MPT case series used for analysis, containing 460 individuals (106 (23%) males and 354 (77%) females) from 440 families is described in Chapter 3. The most frequent tumour types are described in Chapter 3 and Table 3.3 with a more comprehensive list in Table A1. Tumour combination frequencies are described in Chapter 3, Table 3.3 and Figure 3.4.

Prior genetic testing is described in Table 4.4 with reasons for non-detection of the relevant variant illustrated in Figure 4.4. Information regarding previous genetic testing was available for 405/440 (92%) of probands. No molecular investigations had been performed in 91 (20.7%). 159 (36.1%) had undergone *BRCA1*/*BRCA2* testing, 87 (19.8%) had been assessed for Lynch syndrome (where microsatellite instability (MSI) and/or immunohistochemistry (IHC) analysis is considered as assessment) and 159 (20.7%) had had another germline genetic test. The mean number of genes

analysed (where MSI/IHC is considered as analysing four Lynch syndrome genes) was four. Samples from 79 (18%) of probands had undergone sequencing with a multi-gene panel assay with the mean number of genes analysed with these assays being 13.8.

4.1.3.2 - Genetic findings – Single nucleotide variants (SNVs) and indels

Variant filters applied to annotated VCF files produced 89 unique variants in 119 individuals for further ACMG guideline-based assessment. Of these, 22 (42 occurrences) could be classified as pathogenic, 23 (24 occurrences) as likely pathogenic, 24 (27 occurrences) as a variant of uncertain significance (VUS), and 20 (26 occurrences) as likely benign. Six occurrences of P/LP variants occurred in two members of the same family and only three of these contributed to the detection rates quoted below. No pathogenic non-coding variants were identified.

Overall, 63 variants in 17 genes in 61 (13.9%) probands were assessed as P/LP (Table 4.4). Most were nonsense or frameshift variants. Individuals with variants in moderate risk CPGs *CHEK2* (n=14) and *ATM* (n=10) were the most frequent with one homozygote for *CHEK2* ENST00000328354 c.1100delC (p.Thr367Metfs) (annotated in these data as ENST00000382580 c.1229delC (p.Thr410fs)) detected. Individuals with variants in *BRCA2* (n=6), *PALB2* (n=6), *FH* (n=5), *NF1* (n=4), *NTHL1* (homozygous, n=3), *MAX* (n=2), *PTEN* (n=2), *SDHB* (n=2), *BMPR1A* (n=1), *BRCA1* (n=1), *CDKN1B* (n=1), *EXT2* (n=1), *MLH1* (n=1), MSH2 (n=1) and *PMS2* (n=1) were also noted.

Confirmation of the 63 P/LP SNVs/indels detected by WGS was performed by a second analysis (TCP for 52 variants and Sanger sequencing for 11 variants). Pre-testing information was available for 57/63 P/LP variants, 41/57 (71.9%) of which occurred in an individual who had at least one previous genetic test and 7/57 (12.3%) of which were eventually detected by clinical services. No P/LP variants were observed in genes that had previously been tested in the relevant individual by diagnostic services (Figure 4.4). The mean number of genes tested in those with a P/LP variant was 5.3, which was not significantly different to probands without such variants detected (students t-test p=0.396).

Of the 61 probands identified with a P/LP variant, 36 (59%, 8.2% of all probands) had previously been diagnosed with a tumour typically associated with the relevant CPG. A further eight (1.8%) of probands were found to harbour a VUS and had been diagnosed with an associated tumour.

Three probands harboured two P/LP variants in different CPGs. Combinations of variants *PMS2*/*BMPR1A*, *MAX*/*FH* and *FLCN*/*CHEK2* were observed, which are discussed in Chapter 5.

Table 4.4 - Filtered single nucleotide variants and indels deemed pathogenic or likely pathogenic by American College of Medical Genetics criteria

Gene	Transcript	Coordinate	Description	Consequence	Phenotype with age at diagnosis (* indicates tumour deemed typically associated with deleterious variants in gene)	Genes tested by clinical services	Consultation year
ATM	ENST00000278616	chr11:108099912	c.193C>T (p.Gln65*)	Stop gain	NMSC, 14; PNS Nerve sheath benign, 50; Breast, 52[a]; CNS meningioma, 58	PTCH1, NF2 (single gene)	2014
ATM	ENST00000278616	chr11:108175528	c.5623C>T (p.Arg1875*)	Stop gain	Breast, 40[a]; Breast, 45[a]	BRCA1, BRCA2 (excluded in other family members)	2016
ATM	ENST00000278616	chr11:108186841	c.6583+1G>A	Splice site (donor)	NMSC, <40	PTCH1, SUFU (single gene)	2012
ATM	ENST00000278616	chr11:108196843	c.6866-6867insT (p.Ser2289Serfs)	Frameshift	Thyroid, 39; Paraganglioma, 39	SDHAF2, SDHB, SDHC, SDHD, RET, MAX, TMEM127, VHL (panel)	2015
ATM	ENST00000278616	chr11:108115600	c.748C>T (p.Arg250*)	Stop gain	Breast, 48[a]; Colorectal, 60	MSI (stable) BRCA1, BRCA2, MLH1, MSH2 (single gene)	1999
ATM	ENST00000278616	chr11:108205832	c.8147T>C (p.Val2716Ala)	Missense	Breast, 55[a]; Colorectal, 56	No testing	2016
ATM	ENST00000278616	chr11:108214084	c.8405delA (p.Gln2802fs)	Frameshift	Testicular, 21; Thyroid, 35; UKP, 35	No testing	2016
ATM	ENST00000278616	chr11:108180945	c.5821G>C (p.Val1941Leu)	Missense	PNET, 33; Adrenal adenoma, 33	Information unavailable	Unknown
ATM	ENST00000278616	chr11:108205807	c.8122G>A (p.Asp2708Asn)	Missense	Lipoma, <13; Bone benign, <13	Information unavailable	Unknown
ATM	ENST00000278616	chr11:108202751	c.7775C>G (p.Ser2592Cys)	Missense	Hem lymphoid, 9; Breast, 39[a]	No testing	2014
BMPR1A[c]	ENST00000372037	chr10:88676945	c.730C>T (p.Arg244*)	Stop gain	Colorectal, 50[a]; Breast, 57	IHC (PMS2 loss). PMS2 (single gene)	2015
BRCA1	ENST00000471181	chr17:41245586	c.1961-1962insA (p.Lys654fs)	Frameshift	Breast, 38[a]; Haematological lymphoid, 39; NMSC, 56; Ovary, 64[a]	Information unavailable	2014
BRCA2	ENST00000544455	chr13:32913017	c.4525C>T (p.Gln1509*)	Stop gain	Melanoma, 30; Melanoma, 44; Thyroid, 47	No testing	2016
BRCA2	ENST00000544455	chr13:32914174	c.5682C>G (p.Tyr1894*)	Stop gain	PNET, 24; Breast, 40[a]	No testing	2014
BRCA2	ENST00000544455	chr13:32914766	c.6275-6276delTT (p.Leu2092fs)	Frameshift	Thyroid, 38; Colorectal, 57	Information unavailable	Unknown
BRCA2	ENST00000544455	chr13:32914893	c.6402-6406delTAACT (p.Asn2135Leufs)	Frameshift	Testicular, 49; Testicular, 60; Prostate, 68[a]	No testing	2015
BRCA2	ENST00000544455	chr13:32915027	c.6535-6536insA (p.Val2179fs)	Frameshift	Bladder, 53; NMSC, 54; GINET, 55; Aerodigestive tract, 59; Colorectal, 63	No testing	2016

BRCA2	ENST00000544455	chr13:32907420	c.1805-1806insA (p.Gly602fs)	Frameshift	Hem lymphoid, 42; Breast, 43[a]; Endometrium, 49	BRCA2 (not known if single gene or panel)	2016
CDKN1B	ENST00000228872	chr12:12870920	c.148-149delAG (p.Arg50fs)	Frameshift	Paraganglioma, 33; Breast, 34	Illumina TruSight Cancer panel (CDKN1B not included)	Unknown
CHEK2	ENST00000382580	chr22:29091226	c.1392delT (p.Leu464fs)	Frameshift	Kidney, 56; Kidney, 56; Kidney, 56	FLCN, VHL (single gene)	Unknown
CHEK2	ENST00000382580	chr22:29091226	c.1392delT (p.Leu464fs)	Frameshift	Thymus, 53; Breast, 54[a]; Haematological lymphoid, 63; Haematological lymphoid, 67	Information unavailable	2015
CHEK2	ENST00000382580	chr22:29091226	c.1392delT (p.Leu464fs)	Frameshift	Kidney, 56; Kidney, 60	Information unavailable	2010
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Thyroid, 45; Pancreas, 48	No testing	Unknown
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Fibrofolliculoma (multiple), 18; Kidney, 53	FH, FLCN, MET, SDHB, VHL (panel)	2015
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 52[a]; Melanoma, 54	Information unavailable	Unknown
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 50[a]; Kidney, 62; GI NET, 63; Haematological myeloid, 65	MEN1 (single gene). SDHA, SDHAF2, SDHB, SDHC, SDHD, RET, MAX, TMEM127, VHL (panel)	2013
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Endometrium, 54; Breast, 57[a]	IHC (normal), MLH1, MSH2, MSH6 (single gene)	2016
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Kidney, 70; Haematological lymphoid, 70; Colorectal, 72	No testing	2014
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 31[a]; Gastric, 49	BRCA1, BRCA2 (single gene) APC, BMPR1A, CDH1, MLH1, MSH2, MSH6, MUTYH, PMS2, SMAD4, STK11, TP53 (panel)	2015
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 45[a]; Breast, 54[a]; Breast, 55[a]	No testing	2001
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Colorectal, 27; Endometrium, 53; Colorectal, 56; NMSC (multiple), <64	IHC (normal) and MSI (high). MLH1, MSH2, MSH6, PMS2	2016
CHEK2	ENST00000382580	chr22:29105993	c.1051+1C>T	Splice site (donor)	Breast, 46[a]; Ovary, 49; Ovary, 49; Endometrium, 49	BRCA1, BRCA2 (single gene)	2012
CHEK2	ENST00000382580	chr22:29115410	c.784delG (p.Glu262fs)	Frameshift	Colorectal polyps, 46; Parathyroid, 48; Parathyroid, 55; Parathyroid, 59	APC, BMPR1A, CDC73, CDKN1B, MEN1, PKD2, SDHB, SDHC, SDHD, SMAD4, VHL (single gene)	2010
CHEK2	ENST00000382580	chr22:29121242	c.562C>T (p.Arg188Trp)	Missense	Colorectal, 46; Breast, 54[a]; Endometrium, 67	BRCA1, BRCA2, MLH1, MSH2 (single gene)	2007

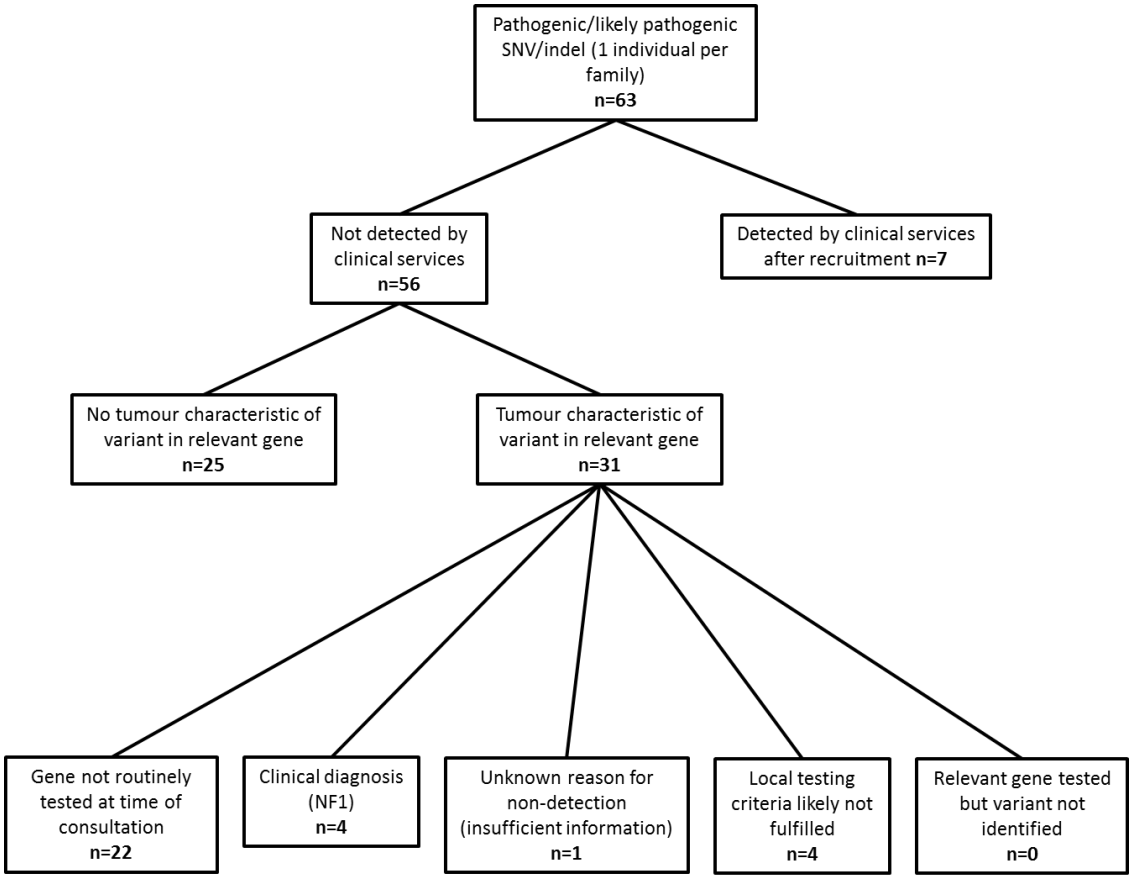
CHEK2[b]	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 40[a]; Pancreas benign, 41	BRCA1, BRCA2, CDH1, CDK4, CDKN2A, MEN1, PTEN, SDHB, STK11, TP53, VHL (panel)	2014
EXT2	ENST00000395673	chr11:44129776	c.613C>T (p.Gln205*)	Stop gain	Breast, 40; Colorectal, 48	BRCA1, BRCA2 (single gene)	2013
FH	ENST00000366560	chr1:241661227	c.1433-1434insAAA (p.Lys477_Asn478insLys)	Inframe insertion	NMSC, 36; Thyroid, 37; NMSC (multiple), 47	Hereditary cancer panel. 24 genes (not specified)	2016
FH	ENST00000366560	chr1:241661227	c.1433-1434insAAA (p.Lys477_Asn478insLys)	Inframe insertion	Small bowel, 53; Colorectal, 56	MSI (stable)	2016
FH	ENST00000366560	chr1:241661227	c.1433-1434insAAA (p.Lys477_Asn478insLys)	Inframe insertion	Breast, 49; Colorectal, 65; NMSC, 65	No testing	2016
FH	ENST00000366560	chr1:241676961	c.320A>C (p.Asn107Thr)	Missense	Cutaneous leiomyoma, 36[a]; Uterine leiomyoma (multiple), 36[a]; Breast, 40	FH (single gene)	2016
FH[d]	ENST00000366560	chr1:241675301	c.521C>G (p.Pro174Arg)	Missense	Phaeochromocytoma, 16; Phaeochromocytoma, 35	SDHB, SDHC, SDHC, RET, VHL (single gene)	2008
MAX	ENST00000358664	chr14:65544637	c.289C>T (p.Gln97*)	Stop gain	Phaeochromocytoma, 31[a]; Phaeochromocytoma, 35[a]	SDHB, SDHC, SDHD, VHL (single gene)	2008
MAX[d]	ENST00000358664	chr14:65569057	c.1A>G (p.Met1Val)	Start loss	Phaeochromocytoma, 16[a]; Phaeochromocytoma, 35[a]	SDHB, SDHC, SDHC, RET, VHL (single gene)	2008
MLH1	ENST00000231790	chr3:37083758	c.1884-1G>A	Splice site (acceptor)	Soft tissue sarcoma, 27; Colorectal, 47[a]	APC, BMPR1A, MLH1, MSH2, MSH6, MUTYH, SMAD4, STK11, TP53 (panel)	2015
MSH2	ENST00000233146	chr2:47690234	c.1452-1455insAATG (p.Leu484-Met485fs)	Frameshift	Breast, 40; NMSC, 40; UKP, 42	BRCA1, BRCA2, TP53, PTEN (panel)	Unknown
NF1	ENST00000358273	chr17:29546035	c.1541-1542delAG (p.Gln514fs)	Frameshift	Nerve sheath benign, <30[a]; GIST, 46[a]; CNS Nerve sheath, 51[a]	No testing	2015
NF1	ENST00000358273	chr17:29588770	c.4620delA (p.Ala1540fs)	Frameshift	Lipoma, 29; GIST, 44[a]	No testing	2015
NF1	ENST00000358273	chr17:29661873	c.5831delT (p.Leu1944fs)	Frameshift	GIST (multiple), 36[a]	No testing	2015
NF1	ENST00000358273	chr17:29684007	c.7768-7769insA (p.His2590fs)	Frameshift	PNS Nerve sheath, 20[a]; GIST, 41[a]	KIT, MAX, PDGFRA, SDHA, SDHB, SDHC, SDHD, TMEM127 (panel)	2016
NTHL1[b]	ENST00000219066	chr16:2096239	c.268C>T (p.Gln90*)	Stop gain	Colorectal, 51[a]; Breast, 57	No testing	Unknown
NTHL1[b]	ENST00000219066	chr16:2096239	c.268C>T (p.Gln90*)	Stop gain	CNS meningioma, 42; CNS meningioma, 42; Colorectal, 44[a]	IHC (normal), MSI (stable)	2015
NTHL1[b]	ENST00000219066	chr16:2096239	c.268C>T (p.Gln90*)	Stop gain	Colorectal, 48[a]; Aerodigestive tract, 50	Information unavailable	2012
PALB2	ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Stop gain	Melanoma, 38; Breast, 47[a]	BRCA1, BRCA2 (single gene)	2011
PALB2	ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Stop gain	Prostate, 71	No testing	Unknown

PALB2	ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Stop gain	Melanoma, 31; Breast, 40[a]	BRCA1, BRCA2 (single gene)	2012
PALB2	ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Stop gain	Anus, 37; Breast, 42[a]	BRCA1, BRCA2 (single gene)	2004
PALB2	ENST00000261584	chr16:23625409	c.3116delA (p.Asn1039fs)	Frameshift	Breast, 35[a]; Skin sarcoma, 37; Aerodigestive tract, 43	BRCA1, BRCA2 (single gene)	2006
PALB2	ENST00000261584	chr16:23649437	c.62T>G (p.Leu21*)	Stop gain	Colorectal, 51; Breast, 54[a]	BRCA1, BRCA2, MUTYH (single gene)	2005
PMS2[c]	ENST00000265849	chr7:6037018	c.741-742insTGAAG (p.Pro247_S248fs)	Frameshift	Colorectal, 50[a]; Breast, 57	IHC (PMS2 loss). PMS2 (single gene)	2015
PTEN	ENST00000371953	chr10:89720852	c.1003C>T (p.Arg335*)	Stop gain	Breast, 35[a]; Ovary, 47; Breast, 49[a]	BRCA1, BRCA2 (single gene)	2010
PTEN	ENST00000371953	chr10:89717672	c.697C>T (p.Arg233*)	Stop gain	Endometrium, 36[a]; Thyroid, 50[a]; CNS meningioma, 59; Kidney, 62	PTEN (single gene)	2016
SDHB	ENST00000375499	chr1:17380442	c.223+1C>A	Splice site (donor)	Paraganglioma, 45[a]; Pancreas, 51	BRCA1, BRCA2, CDH1, CDK4, CDKN2A, CTRC, MAX, NF1, PALB2, PRKAR1A, PTEN, RET, SDHA, SDHAF2, SDHB, SDHC, SDHD, SPINK1, STK11, TMEM127, TP53, VHL (Panel)	2015
SDHB	ENST00000375499	chr1:17349179	c.689G>A (p.Arg230His)	Missense	Paraganglioma, 40[a]; Paraganglioma, 40[a]; Paraganglioma, 40[a]	SDHAF2, SDHB, SDHC, SDHD, RET, MAX, TMEM127, VHL (panel)	2014

List incorporates one individual per family. a - Indicates tumour characteristically associated with pathogenic variant in the relevant gene. b - Homozygous, c - Occurring in same individual. d - Occurring in same individual. All structural variants heterozygous. All coordinates are provided for GRCh37.

UKP - Unknown primary, CNS – Central nervous system, PNS – Peripheral nervous system, NMSC - Non-melanoma skin cancer (includes basal cell carcinoma and squamous cell carcinoma), GI NET - Gastrointestinal neuroendocrine tumour, PNET - Pancreatic neuroendocrine tumour, IHC – Immunohistochemistry, MSI – Microsatellite instability.

Figure 4.4 - Prior genetic testing and reasons for non-detection of pathogenic/likely pathogenic single nucleotide variant or indel



4.1.3.3 - Coverage and comparison with panel

Mean depth in WGS data corresponding to coding bases in the 83 genes analysed was 35X (SD = 7.5) with 100% covered at ≥10X. Coverage was also considered for 68 of these genes that are also sequenced by the TCP assay (Table 4.5). In WGS data 100% of target bases were covered at ≥10X with a mean depth of 35.3X (SD = 7.4). Coverage analysis pertaining to those 68 genes from the 411 (89.3%) participants also undergoing sequencing with the TCP showed 99.1% target bases at ≥10X and a mean depth of 807.3X (SD = 793.2).

Table 4.5 - Genes sequenced by Illumina TruSight Cancer panel that appear on list of 83 analysed genes

AIP	*CDKN2A*	*EXT1*	*MSH6*	*RB1*	*TMEM127*
ALK	*CEBPA*	*EXT2*	*MUTYH*	*RET*	*TP53*
APC	*CHEK2*	*FH*	*NF1*	*RHBDF2*	*TSC1*
ATM	*CYLD*	*FLCN*	*NF2*	*RUNX1*	*TSC2*
BAP1	*DDB2*	*GATA2*	*PALB2*	*SDHAF2*	*VHL*
BMPR1A	*DICER1*	*HNF1A*	*PHOX2B*	*SDHB*	*WT1*
BRCA1	*EGFR*	*KIT*	*PMS2*	*SDHC*	*XPA*
BRCA2	*EPCAM*	*MAX*	*PRKAR1A*	*SDHD*	*XPC*
BRIP1	*ERCC2*	*MEN1*	*PTCH1*	*SMAD4*	
CDC73	*ERCC3*	*MET*	*PTEN*	*SMARCB1*	
CDH1	*ERCC4*	*MLH1*	*RAD51C*	*STK11*	
CDK4	*ERCC5*	*MSH2*	*RAD51D*	*SUFU*	

A comparison of the variant detection rate was performed based on the 105 ACMG assessed SNVs/indels that were detected by WGS and were within a gene sequenced by the TCP. 99/105 variants were called from TCP data with quality indicators sufficient to pass filters used for the WGS data. Five undetected variants were indels where review with IGV showed a VAF below the threshold for filtering, including one P/LP variant in *PMS2* (ENST00000265849 c.741-742insTGAAG (p.Pro247_Ser248fs)) where 58/202 (20.6%) reads contained the insertion. One undetected variant in *TMEM127* (ENST00000258439 c.665C>T (p.Ala222Val)) was covered by only two reads, hence non-detection.

The filtering and assessment process applied to WGS data was also used for variants called from TCP data generated from the same 411 individuals. 108/110 variants from TCP data that passed filters and went forward for ACMG assessment were also called from WGS data, meaning that two variants (assessed as pathogenic) were not detected by WGS. This was due to VAF being marginally below the filtering threshold of 33% for ATM ENST00000278616 c.2426C>A (p.Ser809*) (7/22 (32%) reads) and MAX ENST00000358664 c.97C>T (p.Arg33*) (9/29 (31%) reads).

4.1.3.4 - Comparison of loss of function variant detection rate in Multiple Primary Tumour WGS data and gnomAD dataset

In the MPT dataset, 52 truncating or splice site variants were observed in 440 MPT probands compared with 298 in 8992 gnomAD genomes based on observed variant frequency estimates adjusted to reflect sex distribution of the MPT series (13.6% vs 3.3%, χ^2=84.903 p=<0.0001). 41 truncating or splice site CPG variants occurred in a proband with at least one tumour type uncharacteristic of the relevant CPG and the frequency of such variants in these cases was also

compared to that in gnomAD. This was significantly higher in the MPT probands with truncating/splice site variants and uncharacteristic tumours (41/440 (9.3%) vs 298/8992 (3.3%), χ^2=43.642 P=<0.0001).

4.1.3.5 - Genetic findings – Structural variants

Structural variant analysis revealed seven potentially pathogenic variants in 7/440 (1.6%) probands (Table 4.6), although SV calls were not available for all individuals. Further details of validation of these SVs with Sanger sequencing and IGV plots showing supporting reads (for Manta calls) can be found in Appendix 5 (variants 1-7). Three of these probands had previously been diagnosed with tumours typically associated with variants in the relevant gene with an additional two having a family history of such tumours in a first degree relative (colorectal cancer at age 56 for the case with a *SMAD4* translocation and renal cell carcinoma at age 69 for the case with the *TSC1* duplication). One individual with an inversion of *PTEN* exon 7 had been diagnosed with breast cancer at age 45 and had a strong family history of this tumour, which had occurred in her sister (age 57), mother (age 57), and maternal cousin (age 49). The proband's sister had also been diagnosed with a borderline ovarian mucinous tumour and nasal basal cell carcinoma at ages 46 and 57 respectively but WGS did not detect the *PTEN* inversion in her sample. A further individual had previously been investigated with germline *FH* sequencing following the diagnosis of multiple cutaneous leiomyomas and a family history of a first degree relative undergoing a hysterectomy for uterine leiomyomas. SV analysis revealed whole gene deletion of *FH.* A chromosome 17:10 translocation where the breakpoint was within intron 9-10 of *FLCN* was identified in an individual with fibrofolliculomas and renal cell carcinoma who also carried a truncating *CHEK2* variant (see SNVs and indels results above).

Table 4.6 –Structural variants passing filtering steps

Gene	Chromosome	Predicted start	Predicted end	Algorithm	Predicted consequence following IGV review	Phenotype with age at diagnosis (* indicates tumour deemed typically associated with deleterious variants in gene)	Genes tested by clinical services	Year consulted
FLCN	17	17134310 (Manta), 17134474 (Canvas)	17136696 (Manta), 17137867 (Canvas)	Canvas + Manta	Deletion of exon 2	Breast, 46; Pulmonary lymphangioleiomyomatosis, 47	Information unavailable	Unknown
PTEN	10	89713996	89719837	Manta	Inversion of exon 7	Breast, 45[a]	*BRCA1*, *BRCA2* (single gene)	Unknown
SMAD4	18:9	chr18:48556624.	chr9:127732713	Manta	Translocation with breakpoint within untranslated part of exon 1	CNS, 42 (Colorectal, 56 in mother)	*PMS2, TP53, MLH1* (single gene)	2011
TSC1	9	135803187	135807261	Manta	Duplication of exon 3	Testicular, 47; Prostate, 64; Lung, 70	*BRCA1*, *BRCA2* (single gene Ashkenazi common pathogenic variants)	2016
TSC2	16	1566500	2119769	Manta	Inversion with breakpoint in intron 16-17	Small bowel, 42; Colorectal, 43	IHC (MSH6 loss). *MSH6* (single gene)	2012
FH	1	237244834	242310908	Canvas	Full gene deletion	Multiple cutaneous leiomyomata, <55[a]	FH (single gene)	2014
FLCN	17:10	17:17121531	10:43731507	Manta	Translocation with breakpoint in intron 9-10	Multiple fibrofolliculomas, 18; Kidney, 53.	*FH*, *FLCN*, *MET*, *SDHB*, *VHL* (panel)	2015

List incorporates one individual per family. a - Indicates tumour characteristically associated with pathogenic variant in the relevant gene. CNS – Central nervous system. All structural variants heterozygous.

4.1.3.6 - Combined variant detection rate

If SVs passing filters and ACMG assessed P/LP SNV/indels are combined, a detection rate of 15.2% (67 probands tested) is observed. 39 probands (8.9% of total) had such a variant and a typically associated tumour. There was no significant difference in P/LP detection rate between probands who had been diagnosed with a rare tumour and those who hadn't (24/136 (17.6%) vs 40/304 (13.1%) χ^2=1.5235 p=0.2171). Of the 55/67 probands where family history information was available, there was no cancer diagnosis in a first degree relative under 60 years in 23 cases (41.8%) and under 50 years in 34 cases (61.8%).

Limited numbers of family members participated in the study, preventing large scale segregation analysis. Of the 70 P/LP variants (including SVs) of interest detected in probands, the relevant locus was sequenced in a family member on seven occasions. The relevant variant was detected in 4/7 family members, two of whom had been diagnosed with a typically associated tumour (breast cancer in *PALB2* and *BRCA1* variants).

4.1.4 - Discussion

4.1.4.1 - Variant detection rates in a multiple primary tumour series

A previous retrospective analysis of MPT cases (defined as two primaries under age 60 years) referred to a UK clinical genetics service without prior genetic testing observed that 20.7% (44/212) were found to have a molecular diagnosis upon routine targeted molecular genetic testing including *BRCA1*/*BRCA2*, mismatch repair gene analysis or other single gene testing (*APC*, *MUTYH*, *PTEN*, *TP53* and *RB1*).[196]

In the current study it was considered whether comprehensive genetic analysis in pre-assessed individuals with MPT might increase the diagnostic yield over routine targeted testing. Thus, 460 individuals with MPT were analysed that had previously undergone routine genetic assessment/molecular testing but with no molecular diagnosis made. Interrogation of WGS data for variants in 83 CPGs identified a P/LP variant in 67/440 (15.2%) of probands (incorporating SNVs/indels and SVs), including those affecting moderate and high-risk CPGs.

As the MPT cohort reported here was mostly ascertained from UK genetics centres (and was similar to the cases that were in the previous retrospective cohort that did not have a known genetic cause), it is estimated that (assuming that WGS would detect variants identified by routine targeted sequencing approaches) that comprehensive genetic analysis in a referred series of individuals with MPT with no prior genetic testing would detect a P/LP variant in around a third of cases (20.7% + 12.1% (estimated assuming a diagnostic yield of 15.2% in the 79.3% without a variant on routine testing) = 32.8%). The

estimated proportion of cases with a P/LP variant and a typical tumour would be ~27.5% (20.7% (all of those with detected tested by targeted analysis had a typical tumour) + (79.3% x 8.9% = 7%)). Therefore, in individuals seen in a genetic clinic, the presence of MPT (two tumours below 60 years or three below 70 years) could be taken as an indication for considering genetic testing. These estimates for diagnostic yield are approximate and would be influenced by ascertainment processes but do suggest that comprehensive testing for CPG variants significantly increases the detection of P/LP variants over the targeted testing that has been routinely employed in most genetics centres.

Most MPT cases with a P/LP variant (39/67 (58.2%), 39/440 (8.9%) of all pre-assessed probands tested in the current study) had been diagnosed with a tumour type characteristically associated with variants in the relevant CPG, findings which arguably have greater clinical utility then where no associated neoplasm is seen. In, addition, a further 8/440 (1.8%) had a VUS and a previous diagnosis of a characteristic tumour. Such VUSs might eventually be reclassified as LP variants with further investigations (e.g. tumour studies) or additional clinical information (e.g. segregation analysis). However, interpretation of segregation data should be cautious in cancer predisposition syndromes due to incomplete penetrance and high probability of phenocopies. Tumour studies for loss of heterozygosity do not provide absolute confirmation or exclusion of pathogenicity and together, these considerations reinforce the importance of data sharing initiatives such as ClinVar.[187]

A major influence on the number and pattern of variants detected in a study such as this is the tumour phenotypes occurring in the cohort, which in this case reflect both population incidence and patterns of referral for genetic assessment/investigation (see Chapter 3). Breast cancer accounted for almost a quarter of tumours in the MPT series and most genes in which deleterious variants were detected are breast CPGs, many of which have not been routinely tested in the UK. Pathogenic variants in *ATM* and *CHEK2* are associated with moderate risks[197,198] and these genes had not been tested by the referring centre in any of the cases with P/LP variants. Six probands had pathogenic variants in *PALB2*, a gene initially thought to confer moderate risk[39] but subsequently reported to have a penetrance somewhere between moderate and high risk genes such as *BRCA1* and *BRCA2*.[40]

Genes may remain un-investigated by clinicians not only due to uncertainty surrounding risks but also recency of discovery. A number of CPGs in which variants were identified, such as *MAX* and *FH*, have been relatively recently described. The appearance of these variants in this analysis is likely to reflect lack of availability of testing at the time of consultation and subsequent referral for inclusion in the study. Molecular genetic testing has been available for other genes, such as *MLH1* and *PTEN*, for a greater period of time but some individuals appeared not to have fulfilled the clinical testing criteria applied in the referring centre. For example, an individual with breast and ovarian cancer was identified with a *PTEN* nonsense variant but testing for this gene had not been undertaken by clinical

services. This is presumably either because there was an absence of other manifestations of *PTEN* variants such as macrocephaly, or that they had not been elucidated due to lack of suspicion for that group of disorders. Four individuals were identified with *NF1* P/LP variants and exhibited largely typical neoplastic phenotypes including neurofibromas, malignant peripheral nerve sheath tumour and gastrointestinal stromal tumour. Rather than clinicians not considering the diagnosis, the appearance of these participants amongst the positive results likely indicates that neurofibromatosis type 1 is frequently regarded as a clinical diagnosis where *NF1* sequencing is not required due to reported full penetrance. If practice were to change to a more liberal sequencing approach then it may lead to revision of the natural history of the disease and more data with which to define genotype-phenotype correlations.

TP53 is a further well-established CPG that is associated with diverse and multiple cancers and has clear clinical criteria for testing that are often not fulfilled. Despite this, no pathogenic variants were detected. Germline *TP53* variant related phenotypes (often with rare and/or early onset cancers) are more clearly identifiable clinically and less likely to appear in cohorts such as this without specifically ascertaining for them. Consistent with this are pathogenic variant detection rates of ~4% in earlier onset (≤30 years) breast cancer cases[199] and ~17% in MPT individuals referred for germline *TP53* testing who generally fulfilled criteria for that investigation, had tumours characteristic of Li Fraumeni syndrome and an average age at diagnosis (of a first primary) under 30.[151]

Although this study is, to the author's knowledge, the first report of the application of WGS to an adult MPT series, other studies have used agnostic NGS strategies in single site cancer cohorts. Pathogenic variant detection rate in these analyses may be influenced by the assay used, the variant filtering/assessment applied and the nature of the series in terms of both phenotype and ascertainment. Application of a 76 gene panel to ~1000 adult cancer cases referred for germline genetic testing and ACMG guideline based assessment of resulting variants showed a 17.5% rate,[200] while a similar sized series from the same centre using tumour-normal sequencing in advanced cancer (regardless of genetic testing referral) reported an equivalent figure of 12.6%.[201] The genes containing the most frequent pathogenic variants in both studies are similar to the current study (*BRCA1, BRCA2, CHEK2, and ATM*) but the detection rates are lower than the estimate of around a third of newly referred MPT cases, likely reflecting greater likelihood of a germline pathogenic variant in both genetics referrals and in MPT individuals. Studies of WGS and/or WES applied to unselected paediatric cancer series have also shown pathogenic variant detection rates close to 10% but a contrasting range of affected genes with *TP53* and genes associated with embryonal tumours playing a far greater role.[202–204]

4.1.4.2 - Atypical tumour-variant associations in multiple primary tumour cases

In this study multi-gene testing was applied in all cases irrespective of the tumour types diagnosed. Strikingly, this resulted in the identification of a large number of probands (29/67, 43.2%) who harboured a P/LP CPG variant but whose tumour phenotypes were not entirely typical for the relevant CPG. This situation has been reported at high frequency in other reports of extensive NGS testing of cancer cohorts[200,203,205] and represents a challenge for clinicians because the relevance of the variant to cancer risk in the consultand and their family is less clear. Specific atypical associations observed in this analysis are heterogeneous and numbers are small but some patterns are noted including 5/16 (31.2%) of *CHEK2* variant carriers being previously diagnosed with renal cell carcinoma (RCC) (breast cancer occurred in 8/16 (50%)). An odds ratio of 2.1 for RCC has previously been observed in *CHEK2* variant carriers but only associated with the Ile157Thr founder mutation in a Polish population.[206] 2/6 (33.3%) of *PALB2* variant carriers had cutaneous melanoma under the age of 40 years and 2/10 (20%) individuals with *ATM* variants had thyroid cancer before that age. However, an analysis of 182 melanoma families only demonstrated one pathogenic *PALB2* variant[207] and thyroid malignancies have not been reported at increased frequency in homozygous or heterozygous *ATM* variant carriers.[45,55]

One potential interpretation of atypical tumour phenotypes is that the tumour spectrum associated with some CPGs is wider than currently recognised, in part because to date, testing of particular genes has been limited to specific phenotypes. For example, although *FH* variants were demonstrated to predispose to RCC in 2002, they were only shown to predispose to phaeochromocytoma and paraganglioma 12 years later.[208–210] Therefore, it is suggested that further "agnostic" testing of a comprehensive panel of CPGs in MPT cases could lead to the identification of novel gene-tumour phenotype associations. The observation of a significantly higher rate of loss of function variants associated with non-characteristic tumours in the MPT cohort vs the gnomAD dataset suggests that at least some variants identified in individuals with atypical phenotypes are relevant. However, caution is necessary in automatically linking a pathogenic CPG variant to the observed tumour phenotype without further evidence such as larger studies of variant carriers or tumour studies that demonstrate a causative effect of a variant.

Another possibility is that tumours may occur coincidentally in the presence of a pathogenic constitutional CPG. Variants might be considered causative in some contexts or tissues (therefore likely to pass filtering and assessment) but not in others. For example, an inframe insertion in *FH* (ENST00000366560 c.1433-1434insAAA (p.Lys477_Asn478insLys)) was identified in three cases, none of whom had been diagnosed with typical Hereditary Leiomyoma and Renal Cell Carcinoma tumours. This variant causes recessively inherited fumarate hydratase deficiency and has been

demonstrated to disrupt enzyme activity.[211] However, its significance to cancer predisposition in the heterozygous state is less well defined.

Unusual MPT-CPG associations can occur when an individual harbours variants in multiple CPGs, either due to (at least) one of the variants remaining unidentified through diagnostic testing or because of an interactive effect between them. WGS identified three examples in this cohort. The phenomenon is discussed in Chapter 5 and termed Multiple Inherited Neoplasia Alleles Syndrome (MINAS).[212]

4.1.4.3 - Value of germline WGS in the analysis of multiple primary tumour cases

Although WGS could arguably offer the most sensitive and comprehensive strategy for detecting germline CPG variants, it is resource intensive in terms of sequencing, data storage, and analytical capacity. In this study, conservative variant filtering/assessment and the small number of non-coding variants used for data interrogation reduced the post sequencing burden of variants but small changes to these processes would lead to significant increases with uncertain clinical utility. The approximate WGS cost per sample as part of the BRIDGE project was £1000, consistent with figures collated by the National Human Genome Research Institute in 2016 and higher than the £770 per exome derived from that survey.[213] The TCP assay in the Stratified Medicine Core Laboratory (Department of Medical Genetics, University of Cambridge) is currently charged at around £350 per sample. Justification of the extra costs compared to other NGS assays such as panel tests or WES requires the demonstration that WGS can increase the diagnostic rate over other approaches through enhanced coding SNV/indel detection, SV identification or analysis of non-coding regions.

In this analysis, TCP produced a higher mean depth but slightly lower percentage of target bases covered at ≥10X compared to the equivalent regions in WGS data (99.1% vs 100%). WGS identified one *TMEM127* SNV (assessed as VUS) that wasn't detected by TCP due to the relevant nucleotide being covered by only two reads. There were five additional filtered variants in WGS data that weren't called from panel data, one of which was assessed as likely pathogenic. This was due to the VAF being marginally below the chosen threshold, an issue that also accounted for two pathogenic variants being called from TCP data but not from WGS. Non-detection of lower VAF variants could be resolved through more sensitive bioinformatic filtering of data from either assay. 15 genes on the list of 83 were not targeted by the panel and three pathogenic variants were identified in one of them (*NTHL1*) by WGS. This illustrates the broader scope of WGS but the current results do not suggest that WGS offers greatly enhanced CPG SNV/indel detection at present.

Copy number variation can be detected through read counts in exome or panel data and there are a number of algorithms designed for this task.[38] However, non-uniform coverage can compromise analysis of relative read depth for this purpose and focus on coding regions reduces the chance of

reads covering SV breakpoints. The latter point is particularly pertinent for inversions and translocations. WGS addresses some these issues and identified seven SVs predicted to affect a gene of interest, two of which occurred in an individual with tumours in their personal and family history consistent with variants in that gene. There was no evidence in the medical record of the individual with the *PTEN* inversion exhibiting other clinical features of constitutional variants in this gene but also no record of an examination in a consultation where only *BRCA1*/*BRCA2* testing was anticipated. Whilst the numbers of potentially pertinent SVs are small, these aberrations would unlikely be detected by panel or exome sequencing alone. Copy number variation can be identified from analysis of read counts in WES or panel data[214] but most diagnostic laboratories rely on techniques such as multiplex ligation probe assays (MLPA) to test individual genes. If MLPA is applied to many genes then the cost may make WGS more economical than WES/panel-based testing (with concurrent MLPA) but a detailed cost benefit analysis would be required to investigate this. Furthermore, WGS can detect inversions and translocations that are not characterized by MLPA. A note of caution however, arises from a deletion involving exons 14 to 16 of *BRCA2* that was highlighted by the referring clinician but was not detected through the WGS analysis performed in this study.

Given the current limited benefits of WGS over WES/panel analysis demonstrated in this study, a key advantage of the former approach is the ability to prospectively or retrospectively interrogate regions that are not yet known to be clinically relevant. This includes novel CPGs and it is noted that many of the P/LP variants in this analysis were detected due to the gene/region not being available for testing at the time of consultation. Costs of WGS should therefore be considered in the context of possible future demand for re-investigation and the consequent resource burden required for this if the region of interest (including non-coding) has not been sequenced in the first instance. Adequate systems to prioritise and assess the multitude of non-coding variants generated by WGS for clinical use are not yet in existence.[215] Consequently, few clinically relevant non-coding variants are currently known and none were identified in this analysis. However, evidence of regulatory elements that influence expression of any given gene is accumulating[216] and high throughput functional assays to study them provide the opportunity to define diagnostically significant variants influencing risk of neoplasia.[217] If these processes are successful, the case for WGS as a first line investigative tool would become more compelling.

In summary, this work has demonstrated that the application of comprehensive CPG testing to a cohort of previously investigated MPT cases resulted in the detection of multiple pathogenic variants with relevance to the management of those individuals and their relatives. The finding that comprehensive genetic analysis of MPT cases can frequently result in the identification of pathogenic CPG variants that cannot readily be attributed as causative for the observed MPT clinical phenotype has important implications both for clinical practice and for future research into the phenotypic

consequences of germline CPG variants. Summing together variant detection rates from a previous series of MPT cases ascertained in a similar manner and the present analysis suggests that first-line application of WGS (or other strategies for comprehensive CPG variant detection) to a clinical genetics referral-based cohort of MPT cases would detect a deleterious variant in about a third of cases, a large proportion of which would not have a family history of cancer in a first degree relative.

4.2 Investigation of a clinical scoring system to predict the presence of pathogenic cancer predisposition gene variants in multiple primary tumour cases

4.2.1 - Introduction

Clinical prioritisation strategies guiding genetic testing can be seen as lying along a spectrum where at one end lies the most sensitive approach of testing all individuals who develop a malignancy. At the other more focused end, a more traditional approach of targeting testing to highly suggestive phenotypes exists. Application of germline genetic testing to all cancer patients would produce greater numbers of results with uncertain or limited clinical utility at significant cost and highly targeted testing may produce missed diagnoses while compounding ascertainment biases that influence the phenotypes associated with CPG variants.

An intermediate strategy might be to utilise general indicators of cancer predisposition to prompt agnostic genetic testing and the analysis of MPT cases described in this chapter is illustrative of such an approach. Here, all MPT cases fulfilling the inclusion criteria received WGS but it was postulated that further factors such as total number of tumours occurring in an individual, extent of family history and rarity or estimated heritability of tumours could be incorporated into a scoring system to predict those with P/LP variants within the series. If a scoring system could add specificity and be easily applied in clinical settings, it may inform the diagnostic process undertaken by genetics services.

Therefore, to investigate whether MPT individuals harbouring pathogenic CPG variants could be predicted by clinical indicators, a scoring system was devised, herein referred to as a "multiple tumour score" (MTS). The MTS was based on assigning integer values to each tumour occurring in a single family lineage (including the proband) and taking the sum to produce a single value. A similar, albeit more targeted, system has previously been successfully applied to Hereditary Breast and Ovarian Cancer in the form of the Manchester score.[218] An earlier version of an MTS incorporating age at diagnosis and tumour rarity (Table 4.7) was previously published using data generated from MPT cases referred to clinical genetics services, some of which contributed to the present study. It was shown that around a fifth of individuals who didn't have a molecular diagnosis identified had an

MTS equal to or higher than the median in the diagnosed group, but the predictive capacity wasn't investigated.[196]

Table 4.7 - Previous multiple tumour score[196]

Malignant tumour	Age at diagnosis	Score
Breast, lung, colorectal, prostate, non-melanoma skin, cervical	<30	5
	30–39	4
	40–49	3
	50–59	2
	>59	1
Any other malignant tumour	<50	5
	50–59	3

The original MTS was simple to apply clinically but the grouping of tumour histology/morphology into only two groups led to some high scores awarded to tumours which were unlikely to have had a significant constitutional genetic contribution to their aetiology. For example, cervical cancer has a strong association with human papilloma virus infection and has an incidence peak at a relatively young age. Its grouping with common cancers with generally later onset led to high scores being assigned to earlier onset cervical cancers that were unlikely to reflect higher probability of a cancer predisposition syndrome. Additionally, the chosen integer values to assign to each category in the scoring system were chosen arbitrarily but only one set of values were proposed. In the context of trialling predictive capacity of the MTS, a number of options may reveal a set of preferable values in comparison to others.

In this study therefore, it was aimed to improve the MTS to reflect more factors indicating increased likelihood of tumour predisposition and provide greater differentiation between scores assigned to tumours on the basis of those parameters. The considered variables included age at diagnosis, incidence rate of the tumour and estimated heritability. To assist with constructing a scoring system, an attempt was made to estimate the relative value of scores that should be assigned on the basis of these variables but this did not suggest that it could be estimated with any accuracy. Consequently, a number of different systems were proposed and their ability to predict the presence of a P/LP variant in the MPT series tested through logistic regression analysis. The series was divided into training and test sets with the best performing system from the training set being applied to the test set to assess potential clinical utility.

4.2.2 - Methods

4.2.2.1 - Defining tumours on which to assign scores

Analysis was based on the same 440 MPT probands incorporated in the analysis described in section 4.1. The dependent variable used for logistic regression was the presence or absence of a variant assessed as pathogenic or likely pathogenic by that process (including structural variants), herein referred to as P/LP Var +ve.

Family history was available for 400 probands. Pedigrees and/or other medical records were reviewed in these cases and tumours occurring in a single lineage were recorded in terms of age at diagnosis and tumour type. If two lineages contained tumours to record then the one that would be assigned the highest score according to the original MTS system[196] was used. One intervening relative was permitted between any two members of a lineage.

4.2.2.2 - Individual variables analysis (Script RA4.5)

It was intended that values assigned to tumours in the trialled scoring systems would be weighted to produce higher scores for neoplasms deemed more likely to be due to constitutional genetic predisposition. Whilst age at diagnosis, incidence and heritability are known to be broadly relevant to the probability of a CPG variant being present, a numerical measure of this across cancer types and the relative importance of each factor is not easily arrived at. To attempt to assess this for the purposes of devising scoring systems to apply to a training set, logistic regression analysis was initially performed that separately considered these three factors as independent variables.

In the event of an acceptable fit of the logistic regression models/predictive capacity arising from this process, it was anticipated that the regression coefficients (change in natural log of odds of dependent variable conferred by a unit increase in the explanatory variable) could guide the relative scoring of tumours in the final system/s. For example, if a ten-year decrease in age at diagnosis was associated with the same increase in probability of a pathogenic variant as a 30 percent increase in estimated heritability, the final score increases conferred by both these changes would be equal.

In these initial logistic regressions, values assigned to participants were directly informed by figures relating to these three variables rather than a pre-determined score. Where it was not possible to apply a figure (e.g. no heritability estimate available), these tumours were excluded and participants excluded if this process led to them no longer fulfilling the original recruitment criteria (two primaries before age 60 or three before age 70). This left 370 probands for analysis where 45 individuals were designated as P/LP Var +ve. In this and all further analyses, individual scores where family history wasn't considered were also formulated as availability of family history information was not uniform

and reliability of tumour reporting may vary between cancer type, recruiting centre and family make-up. This allowed the inclusion of 407 probands with 56 P/LP Var +ve individuals.

Designation of independent variable values for probands was undertaken as follows. For age, the mean of age at diagnosis of all tumours counted in a lineage was taken. For incidence, the incidence per 100,000 person-years relevant to each tumour type in a lineage was taken based on Cancer Research UK (CRUK) data[219] and the mean taken. Where incidence figures were not available in CRUK data, the literature was reviewed to obtain them. Estimates are frequently different for males and females and these were considered separately according to the sex of the participant. Many tumours occurring in the series are known to be rare and incidence estimates may be less reliable than for common cancers. Rare cancers can be defined as those with an incidence less than 6 per 100,000 person years.[220] For the purposes of this analysis, any cancer known to be rare and without a reliable incidence estimate was assigned a figure given by the mean incidence of all those in the series with a known incidence lower than 6 per 100,000 person years (1.56 per 100,000 person years for males and 1.91 per 100,000 person years for females). Heritability describes the proportion of variance of a given phenotype attributable to inherited factors. For various cancer types, it has been estimated using statistical techniques that control or adjust for non-inherited factors such as environmental exposure, most notably through twin studies.[49,50] A higher heritability estimate should increase the probability of genetic predisposition contributing to the tumours observed (though this does not imply the relative role of lower vs higher penetrance factors). Therefore, participants were assigned independent variable values based on the mean of percentage heritability estimates of the diagnosed tumours in a lineage. Heritability estimates are not available for a comprehensive range of cancers but a key study of heritability estimates contains a pan-cancer estimate of 33%.[50] This figure was applied to cancers without an estimate unless the population attributable fraction (PAF) of the relevant cancer indicated a lower number. In these cases, a heritability estimate was obtained by 100 – PAF. PAF describes the proportion of variance in the incidence of a cancer attributable to environmental factors. Whilst it is limited by which environmental exposures are measured, high estimates might indicate a more limited role for heritable factors. PAF estimates used here were obtained from CRUK data.[77,221]

Logistic regressions for each variable were performed with the R glm function and goodness of fit assessed with Chi square tests (anova function) where the null hypothesis was an improved model fit with fewer (i.e. zero) independent variables. The pROC package[222] was used to generate receiver operator characteristic (ROC) curves and assess the area under curve for each model.

4.2.2.3 – Assessment of models based on individual variables to inform scoring system (Script RA4.3)

Results from the logistic regressions based on age, heritability and incidence are described in Table 4.8. Outputs with and without consideration of family history are shown. No model was assessed as having a satisfactory goodness of fit as assessed by Chi square tests.

Table 4.8 – Logistic regression outputs based on individual variables

Variable	Family history included	Chi square p value	AUC
Age at diagnosis	Yes	0.1258	0.575
Heredity	Yes	0.1515	0.575
Incidence	Yes	0.3081	0.575
Age at diagnosis	No	0.1575	0.5693
Heredity	No	0.1391	0.5814*
Incidence	No	0.7731	0.5038

*Direction of correlation indicated more heritable tumours reduced probability of pathogenic variant

AUC – Area under curve

4.2.2.4 - Devising a scoring system – Scoring options

Given that there was insufficient evidence to guide relative importance of variables in a scoring system, a range of MTS systems were produced (Table 4.9) to apply to a training set. In order to maximise ease of use in potential clinical settings, the score was integer based and arranged values of the independent variables (age, incidence and heritability) into weighted bands. The incidence bands were designed to reflect the definition of rare tumours then equal gradations up to an incidence level at which the UK top 5 incident cancers are observed (>50 per 100,000 person years). Any tumour falling into a particular band would be scored with the same integer value and the sum of these for the different parameters summed for each tumour. The range of MTS systems proposed were designed to provide different levels of weighting between bands and the previously published system was also applied.[196]

Table 4.9 – Multiple tumour scoring system options

Age band (years)	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6
>59	1	1	1	2	2	1
45-59	2	2	3	4	6	10
30-44	3	4	9	8	18	20
<30	4	8	27	16	54	30
Heritability band (%)	**Option 1**	**Option 2**	**Option 3**	**Option 4**	**Option 5**	**Option 6**
0-25	1	1	1	2	2	1
26-50	2	2	3	4	6	10
51-75	3	4	9	8	18	20
76-100	4	8	27	16	54	30
Incidence band (per 100K person years)	**Option 1**	**Option 2**	**Option 3**	**Option 4**	**Option 5**	**Option 6**
>50	1	1	1	2	2	1
29>50	2	2	3	4	6	10
6.1-28	3	4	9	8	18	20
0-6	4	8	27	16	54	30

4.2.2.5 - Assigning scores – Scoring systems (Script RA4.3)

Occurrent tumours in probands and their relatives in a single lineage were each assigned scores according to the proposed systems. Tumours occurring at distant locations or in the same organ pair (in the same individual) received separate scores. If it was evident that distinct multiple tumours had occurred in the same organ (e.g. skin) then scoring was applied as for two tumours. For cancers of unknown primary site, the lowest score possible for a tumour diagnosed at the relevant age was assigned. If age at diagnosis was unknown the oldest age band was assumed.

Applications of the scoring systems were undertaken that both incorporated and ignored the incidence component due to the fact that the most frequently diagnosed cancer predisposition syndromes cause common tumour types and many common tumours have a high estimated heritability.[50] As previously, analysis was also performed with and without consideration of family history. Where family history was considered, 400 probands were included of which 54 were P/LP Var +ve. Where family history was not considered, 440 probands were included incorporating 66 P/LP Var +ve individuals.

The data (with and without family history) were split into training and test sets of equal size based on random designation of P/LP Var +ve cases to each group and a separate randomisation of cases without pathogenic variants (R sample function). Logistic regression for each scoring system was then performed as above. If a score within the system could not be assigned to a tumour (e.g. no heritability band for benign tumours due to no available estimate) then that tumour was not added to the lineage score. This did not result in any exclusion of probands due to insufficient qualifying tumours to fulfil the original recruitment criteria. Assessment of models and their predictive capacity

incorporated area under ROC curve, chi square goodness of fit tests and consideration of whether a higher score led to an increase or decrease in the probability of an individual being labelled as P/LP Var +ve.

4.2.3 - Results

Performance of the models on the training set are shown in Table 4.10. All but one model had a goodness of fit insufficient to produce a Chi square p-value below 0.05 or an area under curve (AUC) suggestive of good predictive capacity. The best performing score where family history was incorporated was system 3 without the incidence component (Chi square p=0.1118, AUC 0.6158). System 3 (with incidence component) performed best in those assessments where family history was not incorporated (Chi square p=0.03451, AUC 0.5954).

Table 4.10 - Training set model outputs ordered by area under curve

Scoring system	Family history incorporated	Incidence component incorporated	Chi square p value	Area under curve
3	Yes	Yes	0.190	0.619
3	Yes	No	0.112	0.616
3	No	Yes	0.035	0.595
5	No	No	0.103	0.589
2	Yes	Yes	0.345	0.581
2	Yes	No	0.244	0.575
5	Yes	No	0.231	0.572
2	No	Yes	0.113	0.569
6	No	No	0.163	0.567
4	Yes	No	0.387	0.554
Original MTS	No	Yes	0.300	0.554
1	Yes	No	0.384	0.545
5	No	Yes	0.945	0.544
6	Yes	No	0.521	0.543
1	Yes	Yes	0.536	0.540
Original MTS	Yes	Yes	0.706	0.539
1	No	Yes	0.387	0.538
5	Yes	Yes	0.503	0.529
4	Yes	Yes	0.629	0.523
2	No	No	0.339	0.523
1	No	No	0.521	0.517
4	No	Yes	0.803	0.513
6	Yes	Yes	0.724	0.511
6	No	Yes	0.838	0.499
3	No	No	0.250	0.459
4	No	No	0.133	0.427

These two models were then applied to the test set (Table 4.11) with family history incorporated (system 3 without incidence component applied, Figure 4.5) and without family history incorporated (system 3 applied). Goodness of fit was poor in both cases and predictive capacities showed little evidence of clinical utility with AUCs of 0.6301 and 0.5309 for system 3 without incidence component and system 3 respectively. It was considered what these sensitivities and specificities might mean if applied in clinical settings. Scores and P/LP Var +ve status were manually reviewed in the test sets to locate a hypothetical optimum score cut-off that would guide whether to perform genetic testing or not. For the test set incorporating family history, a cut-off of 28 would save performing 75/177 (42.4%) tests but miss 4/25 (16%) pathogenic or likely pathogenic variants. For the test set without family history a cut-off of 24 would save performing 54/198 (27.2%) tests but miss 8/28 (28.6%) molecular diagnoses.

The best performing model was also applied to a further test set comprised of 212 individuals (44 P/LP Var +ve) from the series described in the publication where the original MTS system was devised.[196] Unlike the current MPT cohort, these cases were not ascertained to have no identified causative CPG variants despite clinical assessment. No family history was recorded in this series so only scoring system 3 was applied.

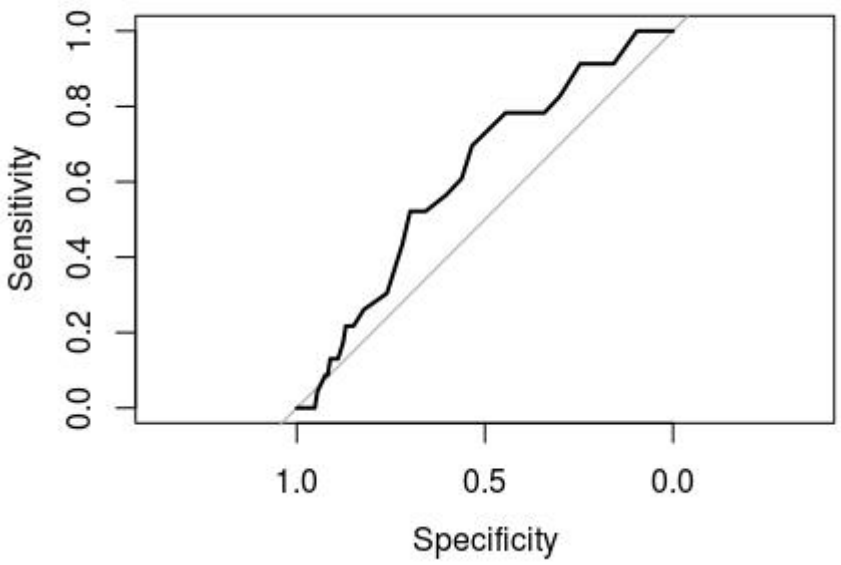
The goodness of fit assessment produced a Chi square p-value (0.06002) that was not significant at a threshold of 0.05 but lower than for other logistic regressions applied. The AUC was 0.6216, which was the highest observed value in these analyses. At a hypothetical MTS cut-off of 20 (considered to be the optimum from manual inspection of the results), application of this system to this series to guide clinical genetic testing would result in 61/212 (28.8%) of individuals not undergoing testing with an associated cost of 4/44 (9%) missed P/LP variants.

Table 4.11 - Application of best performing models to test sets

Test set	Scoring system	Family history incorporated	Incidence component incorporated	Chi square p value	AUC
MPT individuals from present analysis	3	No	Yes	0.483	0.531
MPT individuals from present analysis	3	Yes	No	0.229	0.630
212 MPT individuals from previous study	3	No	Yes	0.060	0.622

AUC – Area under curve

Figure 4.5 - Receiver operator characteristic curve for scoring system 3 without incidence component (on test set incorporating family history). Plot shows result from later randomisation of training and test sets with area under curve of 0.62.



4.2.4 - Discussion

To attempt to produce a scoring system that could predict the presence of a pathogenic variant in an MPT case series, MTS systems were devised and applied to individuals included in the comprehensive CPG analysis described in section 4.1. High penetrance cancer predisposition syndromes are rare disorders conferring significant risk to affected individuals. Therefore, sensitivity is paramount in diagnostic assays designed to detect them. Although a degree of predictive capability for some of the MTS systems was suggested, none performed sufficiently well to suggest an adequate balance of sensitivity and specificity.

Of note is the fact that the MPT WGS series to which the scoring systems were applied was pre-assessed before recruitment to the study and any individuals identified with pathogenic CPG variants by clinical services would not have been invited. MPT individuals diagnosed in the clinic could potentially have phenotypes and family histories that are more obviously indicative of cancer predisposition, leading to higher scores following application of MTS. This non-ascertainment of clinically diagnosed individuals may have led to the P/LP Var +ve group not being adequately representative of unselected cases or sufficiently differentiated from the P/LP Var -ve group to beget good performance of models when applied to the training set. A potential way to address this issue would be to include the 44 P/LP Var +ve individuals from the previously published unselected series in the training set but this would only be applicable to scoring systems that didn't incorporate family history as information regarding tumours in relatives was not recorded for those individuals.

Adaptations to the scoring system may also yield a better correlation between score and pathogenic variant status. One difficulty with the age at diagnosis component is that although cancer becomes more common with age, incidence of individual cancer types does not have a uniform distribution. For example, testicular cancer has a peak incidence between the ages of 30 and 34 and cervical cancer has a bimodal incidence peak.[177] Even cancer types conforming to typical age distribution patterns have varying proportions of cases diagnosed at particular ages. A standard age weighting for all tumours may therefore not reflect likelihood of an inherited cancer syndrome. Age scores more specific to each tumour type may be of benefit but this would add significant complexity and compromise ease of use in the clinic.

Ultimately the central issue in attempting to produce a scoring system to predict the presence of any pathogenic CPG variant may be that cancer predisposition syndromes behave differently. Attempting to identify them all based on a simple scoring system may fail to allow for this complexity and will inevitably predict variants in some genes better than others. For example, in these models a syndrome strongly predisposing to tumours in middle age is likely to produce lower scores than an equally penetrant condition causing susceptibility in younger age groups. Success in predictive models in cancer genetics has tended to centre on using syndrome specific indicators to predict presence of a deleterious variant. Such indicators have been based on relatively well characterised cohorts where extensive details such as histological subtype can be elucidated. The phenotype of cancer predisposition syndromes as an entity *per se* may not be sufficiently well defined at present for this kind of scoring system to be effective.

4.3 Interrogation of cancer panel data for possible clinically relevant mosaic variants

4.3.1 - Introduction

Mosaicism refers to the situation where an individual is composed of two or more genetically distinct cell lines due to early postzygotic genetic changes.[223] This appears to be a frequent phenomenon, potentially affecting a wide variety of loci.[224] Cancer susceptibility may result from mosaicism for a variant in a CPG and this phenomenon is well recognised as a cause of tumour predisposition that may evade detection by conventional genetic testing. Neurofibromatosis type 2 is a condition associated with various central nervous system tumours, particularly vestibular schwannomas. It is caused by pathogenic variants in the *NF2* gene and mosaicism for a cell population containing them is estimated to account for around a third of cases.[225] A recent study of 108 individuals with phenotypes suggestive of Li Fraumeni syndrome identified six mosaic *TP53* pathogenic variants using high depth sequencing[226] and a case of bilateral breast cancer due to a mosaic *BRCA1* exon deletion has been reported.[227]

Mosaicism has significant implications aside from influencing variant detection in the laboratory. It can lead to attenuated phenotypes or be associated with a lack of family history that may prevent further investigation for the condition in question. When detected, it is of reassurance to other family members as mosaic variants are not inherited (notwithstanding the possibility of germline mosaicism where the cell population with the variant is present in ovaries or testes).

Cell populations containing deleterious variants in CPGs may not be represented in blood and present obvious difficulties with detection, even with NGS techniques. More examples of this situation are emerging such as the finding of identical *HIF2A* variants in a patient's paraganglioma and somatostatinoma that explained both tumour's formation. The variant was not detected, however, in blood or other samples including urine, buccal cells and nails.[228] In the not uncommon scenario where multiple tumours occur in the same patient,[111] it may be advantageous to perform genetic analysis on both tumours. Such analysis may become more widespread as NGS technologies are applied in surgical and oncological settings more frequently.

The detection of mosaicism by blood sampling depends on variant carrying cells making up at least a proportion of circulating nucleated cells. If this is the case, the probability of detecting them will be enhanced by a greater number of distinguishable molecular enquiries in the analysed DNA sample for a given genomic coordinate of interest. Chromatogram peaks from Sanger sequencing visually represent the relative proportions of bases at a particular position. They may reveal mosaicism but do not give a quantified measurement of read depth or VAF and suggestive chromatogram profiles may be easily interpreted as artefact. NGS techniques are also imperfect for detection of mosacism but are

often more sensitive for this purpose due to their ability to quantify a particular base call in hundreds or thousands of individual reads, revealing variants that are present in only a small proportion of cells from which DNA was extracted. As per Sanger sequencing however, these reads may be interpreted as artefact and bioinformatic processes are more likely to detect true mosaic variants if optimised for that purpose.

4.3.2 - Methods

To investigate whether mosaic variants in CPGs (detectable in blood) could explain some MPT cases, sequence data from TCP was analysed. This assay is more suited to this purpose than WGS due to the higher read depth (see section 4.1).

4.3.2.1 - Selection of genes and participants

CPGs selected to investigate (n=61, Table 4.12) were those appearing in the gene list for WGS-based comprehensive CPG analysis that are also sequenced by the TCP. CPGs only associated with recessive cancer predisposition were excluded as mosaicism for homozygous/compound heterozygous pathogenic variants in the same gene due to post zygotic mutation is a highly unlikely scenario. Furthermore, mosaicism for monoallelic variants would not be readily distinguishable from biallelic with the sequencing technique utilised.

Table 4.12 - Genes investigated for possible mosaic variants

AIP	*CDK4*	*FH*	*NF1*	*RET*	*TMEM127*
ALK	*CDKN2A*	*FLCN*	*NF2*	*RHBDF2*	*TP53*
APC	*CEBPA*	*GATA2*	*PALB2*	*RUNX1*	*TSC1*
ATM	*CHEK2*	*HNF1A*	*PHOX2B*	*SDHAF2*	*TSC2*
BAP1	*CYLD*	*KIT*	*PMS2*	*SDHB*	*VHL*
BMPR1A	*DDB2*	*MAX*	*PRKAR1A*	*SDHC*	*WT1*
BRCA1	*DICER1*	*MEN1*	*PTCH1*	*SDHD*	
BRCA2	*EGFR*	*MET*	*PTEN*	*SMAD4*	
BRIP1	*EPCAM*	*MLH1*	*RAD51C*	*SMARCB1*	
CDC73	*EXT1*	*MSH2*	*RAD51D*	*STK11*	
CDH1	*EXT2*	*MSH6*	*RB1*	*SUFU*	

The considered MPT cases (n=549) comprised those probands appearing in the WGS-based comprehensive CPG analysis who also had TCP performed on their sample (n=410). 129 other probands were also included who fulfilled eligibility criteria to be included in that analysis but where WGS had not been carried out. An additional 10 individuals were added whose eligibility was dependent on considering multiple (≥10) colorectal polyps as a qualifying tumour.

4.3.2.2 - Bioinformatic processing and filtering (Script RA4.6)

BAM files generated from TCP sequencing output were subject to variant calling as described previously but aligned to hg38. Resulting VCF files were annotated with Annovar.[229] Output files included a measure of VAF. Variant calling was set up to allow heterozygous calls even with a low VAF.

Variants were filtered according to the following criteria: 1) Occurring within a region corresponding to a list of canonical transcripts generated from the gene list (Ensembl transcript identifier converted to RefSeq[230] with Biomart[185]), 2) Read depth ≥200, 3) VAF between 0.05 and 0.3, 4) Allele frequency in 1000 Genomes data (all populations) < 0.01, 5) no indication of a variant call due to multi-mapped reads. Multi-mapping describes a situation where sequencing reads align to more than one region of a reference genome due to sequence similarity between those regions. A read sequenced from a part of the genome with similarity to a region of interest (e.g. a pseudogene) may contribute to variant calls pertaining to the region of interest as it is likely that the two locations will not have identical sequence. This is particularly relevant to variants with low VAFs and the variant calling/annotation pipeline used here included an assessment of the proportion of reads used for that variant call that also aligned to another genomic location. No variants with a proportion above 10% appeared in the annotation output files and only variants with a proportion of 0% were used in this analysis.

Filtered variants were considered for further assessment if the predicted consequence indicated protein truncation ("stop_gain" was the only such annotation in the filtered variants), if there was evidence of pathogenicity in ClinVar[187] (≥2* evidence of pathogenic or likely pathogenic effect corresponding to multiple submissions with no conflicts as to assertion of clinical significance), or if the variant was assigned a DM status in HGMD.[188] An in-house tool to provide a numerical assessment of likelihood of functional alteration using an amalgamation of various in silico tool outputs (unpublished) was also applied to variants. Variants could also be considered further by the designation of a score suggesting a high probability of a deleterious effect (threshold 0.75 on a scale of 0 to 1 where 0 indicates low probability).

Highlighted variants were subsequently reviewed with IGV to check for sequencing artefact. In the majority of cases, all bases contributing to the variant call were in an identical position within read ends. Variants exhibiting this pattern were excluded.

4.3.2.3 - Calculation of coverage (Script RA4.6)

For BAM files from panel data, coverage statistics for regions of interest were generated with samtools depth.[170] A BED file compiled using Ensembl BioMart[185] to represent coding bases of the 61

genes considered was utilised. Mean depth, standard deviation and percentage of target bases covered at a specified depth were calculated using R version 3.4.3.[178]

4.3.3 - Results

The mean sequencing depth across considered coding bases was 796.6X (SD 795.3). 84.4% of bases were covered at sufficient depth to pass the depth filter.

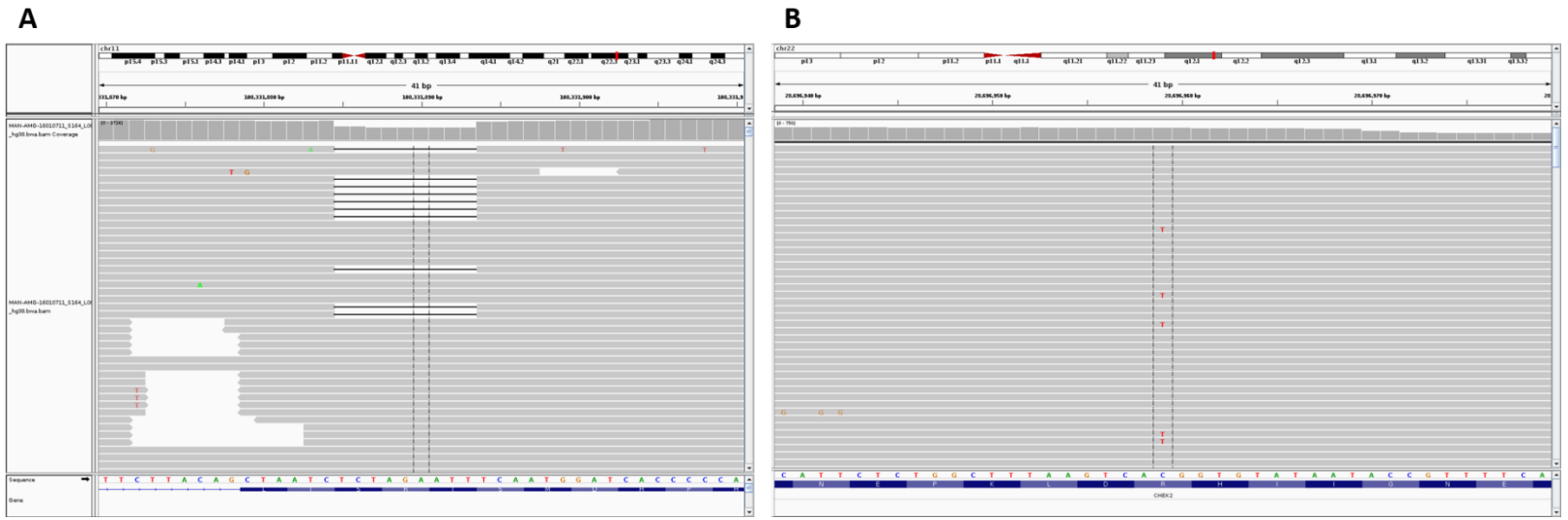
Two variants passed filters (Table 4.13) and were assessed with ACMG criteria in the same manner as those resulting from the WGS analysis described in section 4.1.

Table 4.13 - Variants passing filters to elucidate mosaic variants

Gene	Consequence	Transcript	Variant	Variant allele fraction	Sequencing depth	Phenotype	Family history
ATM	Inframe deletion	ENST00000278616	c.7638_7646delTAGAATTTC (p.Arg2547_Ser2549del)	0.27	3354	AML, 12; Breast, 28	Father - Prostate, 56; Paternal uncle, UKP, 69; Paternal grandfather, Prostate, 50-59
CHEK2	Missense	ENST00000382580	c.1166G>A (p.Arg389His)	0.10	436	Hodgkin lymphoma, 17; Breast, 52; Papillary thyroid carcinoma, 52; Haemangioma (pelvic bone), <54	Paternal grandmother - Oral cancer, 65; Paternal great uncle - Throat cancer, 53; Paternal great uncle - Lung, 67; Paternal great aunt - Breast, 50-59; Paternal great aunt - Breast, ? age; Paternal great aunt - UKP, 20-29; Paternal great uncle - UKP, ? age

AML – Acute myeloid leukaemia, UKP – Unknown primary

Figure 4.6 - A) ATM c.7638_7646delTAGAATTTC (p.Arg2547_Ser2549del). Variant allele fraction 0.27. B) CHEK2 c.1166G>A (p.Arg389His). Variant allele fraction 0.1



ATM ENST00000278616 c.7638_7646delTAGAATTTC (p.Arg2547_Ser2549del) was identified at a variant allele fraction of 0.27 (Figure 4.6) in an individual with childhood acute myeloid leukaemia and subsequent breast cancer at the age of 28, the latter of which is consistent with constitutional pathogenic variants in *ATM*. There was some family history of prostate cancer but no breast cancer was reported in relatives. The variant was assessed as likely pathogenic due to its nature as an inframe deletion, multiple reports of pathogenicity in ClinVar (nine pathogenic and one VUS reports) and published functional evidence of absent kinase activity following transfection into an *ATM* null cell line.[231]

A further *CHEK2* missense variant c.1166G>A (p.Arg389His) at VAF 0.1 (Figure 4.6) passed filters due to predicted high probability of pathogenicity by an in-house in silico prediction tool. The variant was identified in an individual whose various diagnosed tumours included breast cancer but assessment designated it as a VUS. Six reports exist in ClinVar, all with VUS assertion.

4.3.4 - Discussion

Interrogation of CPG variants called from panel data for possible mosaicism resulted in only one variant that was assessed as pathogenic or likely pathogenic. The low number may, in part, be due to inadequate sequencing coverage in some areas. Although the mean depth across considered coding bases was 796.6X (SD 795.3), 15.6% of bases were represented by fewer than 200 reads, the selected threshold for filtering.

The likely pathogenic inframe deletion in *ATM* was identified in a sample from an individual who had previously been diagnosed with early onset breast cancer, suggesting a possible role in causing the tumour phenotype. Tumour material was not available for further investigation in the form of loss of heterozygosity analysis. In theory, mosaic pathogenic CPG variants due to postzygotic mutation shouldn't be associated with a significant family history of neoplasia. In this case, prostate cancers occurring in the father and grandfather at the relatively early age of 50-59 might suggest some constitutional genetic cancer predisposition in that lineage but prostate cancer is not associated with *ATM* variants. Conclusions, therefore, can't be drawn as to the significance of the family history.

Of note is that this individual was diagnosed with acute myeloid leukaemia (AML) at age 12 years, which is generally treated with chemotherapy regimens. They were treated at a time before bone marrow transplant was widely practiced so the variant is unlikely to be derived from a bone marrow donor. Cancer risks are reported to be increased in survivors of childhood cancer survivors[232] and a study of 501 childhood AML cases demonstrated a standardised incidence ratio for any cancer of 10.64, although no breast cancers were noted amongst only five reported second malignancies.[233] Clear associations between treatment and later tumours are difficult to firmly establish (see Chapter 1)

but a role for chemotherapy in causing the breast cancer appears a strong possibility. Chemotherapy may have acted in conjunction with the *ATM* variant as it may have led to a compromised response to DNA damage caused by the drug regimen and an increased rate of tumourigenic events in cells. Alternatively, chemotherapeutic agents may have caused the inframe deletion in a clone of cells. Low VAF variants have been demonstrated in relapsed AML patients, the pattern of which is influenced by the drugs that are used for the initial therapy.[234] However, non-blood cells (such as those in ductal breast tissue) were not considered by the study.

The VAF in this case was relatively high (0.27), increasing the probability that this individual is, in fact, germline heterozygous for the variant. Analysis of WGS data for clinically relevant variants described in section 4.1 demonstrated a number of variants where one assay produced a VAF leading to confident heterozygous designation but another gave a value that fell below the threshold for this assertion. A further sequencing assay was not performed for the individual with the *ATM* variant but this may have shown a higher VAF. When sampling blood DNA, uncertainty as to whether a low VAF for a variant indicates mosaicism in other tissues makes alternative sampling strategies more compelling. A more direct measurement of mosaicism for CPG variants causing multiple primaries is the demonstration of a particular variant in more than one tumour sample but absence in other non-tumour samples (e.g. blood), a phenomenon that has been observed previously.[228] A mosaic variant might also be revealed by absence in blood but presence in a single tumour in which evidence of a second "hit" exists (e.g. two deleterious single nucleotide variants or a single variant with no heterozygosity observed at that locus) because the presence of two mutational events at the same locus can imply that one of them occurred at an early embryological juncture. This rationale has been used in the diagnosis of mosaic NF2[235] but sequencing of a second tumour may be required to identify which of the "hits" is mosaic and which has occurred only in the tumour at hand. The extensive acquisition and sequencing of tumour samples from MPT individuals may yield more positive results than the present analysis but present challenges if formalin fixed paraffin embedded (FFPE) tissue is used due to degradation of DNA stored in that form. Fresh frozen tissue is better suited to sequencing studies but requires prospective organisation of acquisition and changes in routine pathology laboratory practice.

The paucity of possible pathogenic mosaic variants proposed by this analysis may be simply due to the fact that it is a rare phenomenon that has not been widely reported outside of the context of a few conditions. The high rate of mosaic *TP53* variants in phenotypes suggestive of Li-Fraumeni syndrome may be due to the fact that a more specific phenotype was considered where variants in a particular gene are more likely. Some multiple primary tumours caused by the same mosaic CPG variant might be explained by variants that are incompatible with life in the heterozygous state. Genes containing such variants would not be readily identifiable as CPGs in research studies due to lack of surviving

affected individuals and would not have been considered here. The sequencing depth of WGS generated as part of this project would be inadequate to confidently call mosaic variants in putative "mosaic only" CPGs but future studies involving broad coverage of genomic regions with higher sequencing depth might be rewarding in this regard.

Chapter 5 – Multiple Inherited Neoplasia Alleles syndrome (MINAS) – The occurrence of more than one pathogenic cancer predisposition gene variant in the same individual

This chapter is based on, and expanded from, a previously published journal article (Whitworth et al).[212]

5.1 - Introduction

In clinical practice the maxim of Occum's razor is often adopted[236] in the sense that whenever possible, a single diagnosis is favoured over multiple diagnoses. Rare diseases have a frequency of less than one in 2000[237] and statistically, the chances of an individual being affected by two or more of them would appear to be remote. However, with more than 6,000 rare diseases and up to 6-8% of the European population estimated to have such a condition at some point in their lifetime,[237] there is clearly potential for two or more rare disorders to occur by chance. This scenario has been reported in various constitutional genetic disorders with both distinct and overlapping phenotypes, including high penetrance cancer predisposition syndromes and/or patients with multiple primary tumours. If Occum's razor is applied then the detection of a pathogenic variant in a specific cancer predisposition gene (CPG) might lead the clinician to attribute any tumours that are not typical features of the relevant inherited cancer syndrome to variable phenotypic expression or coincidence. In such circumstances, the patient may receive suboptimal management and the estimated cancer risks to relatives could be erroneous. In addition, studies of patients harbouring multiple deleterious variants in different CPGs could provide insights into how the function of the relevant gene products may be related e.g. if a particular combination resulted in a more pronounced or novel phenotype (analogous to the differences in phenotype between patients with monoallelic and biallelic mismatch repair (MMR) gene variants[238]).The best known examples of patients with multiple CPG aberrations are reports of patients with pathogenic variants in both *BRCA1* and *BRCA2.*[239–257] Interestingly, the phenotype in these patients has generally not been shown to be more severe than when a single variant is present.

Through studies undertaken in the author's laboratory to elucidate the constitutional genetic basis of suspected cancer predisposition, ten further individuals (from nine families) have been identified with multiple pathogenic CPG variants that would in themselves be considered to confer sufficient risk to prompt mitigation strategies. Three of these were detected as part of the whole genome sequencing (WGS) based comprehensive CPG analysis in multiple primary tumour (MPT) cases described in Chapter 4 and involved combinations of variants in *BMPR1A*/*PMS2*, *FH*/*MAX* and *CHEK2*/*FLCN* (translocation). Other studies showed combinations of variants in *FLCN*/*NF1*, *FLCN*/*TP53*, *TP53*/*MSH2*, *MLH1*/*XPA*, *NF1*/*BRCA2* and *SDHA*/*PALB2* in individuals with various neoplastic phenotypes.

To provide a summary of the nature and frequency of similar cases reported to date, the published literature was reviewed in systematic fashion. The term "Multilocus Inherited Neoplasia Alleles

Syndrome" (MINAS) is proposed to describe this phenomenon in order to assist with sharing of information regarding the phenotypic effects of particular variant combinations.[212]

5.2 - Methods

5.2.1 - Identification of cases in the literature

In order to review published cases with MINAS, a systematic review of the published literature was undertaken. Initially, a list of CPGs (Table 5.1, n=109) was constructed comprising all genes sequenced by the Illumina TruSight Cancer panel (Illumina Inc., San Diego, CA, USA) and those used for the comprehensive CPG analysis (Chapter 4) that are not targeted by that assay.

The list was then used to perform a Medline database search (1946 to present). Firstly, each gene was entered as a search term (if in existence in the database) and a keyword to produce a list of articles pertinent to that gene. Secondly, the entries were combined with the OR operator to produce 109 lists, each of which contained the articles pertinent to all the genes except one. Thirdly, each of the original individual gene entries was combined via the AND operator with the combination entry that lacked that particular gene. Therefore, articles referring to a given gene name in combination with any other CPG from the list would be captured. Finally, the resulting lists were combined to produce a single entry, which was further combined via the AND operator with the linked terms/keywords "germline mutation" OR "germline" OR "germ-line" OR "double heterozygosity" OR "double heterozygote" OR "genetic predisposition to disease" OR "inherited mutation". An additional PubMed search was also performed using the search term "double heterozygote + cancer."

Titles or abstracts from resulting articles were read to assess whether they reported a case of MINAS and variants described were subsequently reviewed to assess pathogenicity. Variants (and consequently cases harbouring them) were included if it was asserted by the publication that they were pathogenic and there was a predicted truncating consequence (unless benign status in ClinVar), there was pathogenic/likely pathogenic status in ClinVar (2* or 3* evidence unless otherwise stated below) or if they are used in current clinical guidelines to predict increased risk. Variants could also be designated as pathogenic if the article included studies (e.g. reverse transcriptase polymerase chain reaction) that demonstrated abnormal splicing resulting from the variant. It has been speculated that lower penetrance variants may confer increased phenotypic severity when in combination with pathogenic changes in another gene but these cases were not considered.

Table 5.1: Genes used for literature search (n=109)

AIP	*CEP57*	*FANCF*	*NBN*	*RECQL4*	*TMEM127*
ALK	*CHEK2*	*FANCG*	*NF1*	*RET*	*TP53*
APC	*CYLD*	*FANCI*	*NF2*	*RHBDF2*	*TSC1*
ATM	*DDB2*	*FANCL*	*NSD1*	*RUNX1*	*TSC2*
AXIN2	*DICER1*	*FANCM*	*NTHL1*	*SBDS*	*VHL*
BAP1	*DIS3L2*	*FH*	*PALB2*	*SDHA*	*WRN*
BLM	*EGFR*	*FLCN*	*PDGFRA*	*SDHAF2*	*WT1*
BMPR1A	*EPCAM*	*GATA2*	*PHOX2B*	*SDHB*	*XPA*
BRCA1	*ERCC2*	*GPC3*	*PMS1*	*SDHC*	*XPC*
BRCA2	*ERCC3*	*HFE*	*PMS2*	*SDHD*	
BRIP1	*ERCC4*	*HNF1A*	*POLD1*	*SERPINA1*	
BUB1B	*ERCC5*	*HRAS*	*POLE*	*SLX4*	
CDC73	*EXT1*	*KIT*	*POLH*	*SMAD4*	
CDH1	*EXT2*	*MAX*	*PRF1*	*SMARCA4*	
CDK4	*EZH2*	*MEN1*	*PRKAR1A*	*SMARCB1*	
CDKN1B	*FANCA*	*MET*	*PTCH1*	*SMARCE1*	
CDKN1C	*FANCB*	*MLH1*	*PTEN*	*SRY*	
CDKN2A	*FANCC*	*MSH2*	*RAD51C*	*STK11*	
CDKN2B	*FANCD2*	*MSH6*	*RAD51D*	*SUFU*	
CEBPA	*FANCE*	*MUTYH*	*RB1*	*TGFBR1*	

5.2.2 - Tumour studies (for *PALB2*/*SDHA* variants)

Demonstration of loss of the wild type allele in DNA samples obtained from tumours can indicate that a "second hit" occurred at the locus containing a constitutional variant, providing evidence that the constitutional variant was significant in the development of that tumour. For two cases (a mother and son diad), loss of heterozygosity (LOH) analysis was performed for *SDHA* (panel-based sequencing) and *PALB2* (Sanger sequencing).

5.2.2.1 - DNA extraction from formalin fixed paraffin embedded tumour blocks

Slides were prepared from formalin fixed paraffin embedded (FFPE) tumour blocks by the Human Research Tissue Bank, Cambridge University Hospitals. De-paraffinisation was performed by soaking in 100% xylene, 100% ethanol and air drying. In order to optimise the amount of tumour material contributing to sequencing results, slides were reviewed by a pathologist to mark selected tissue and tumour dissection was performed by colleagues in the Department of Haematology and Oncology diagnostic services, Cambridge University Hospitals. Resulting tissue was placed in ATL tissue lysis buffer (Qiagen, Hilden, Germany) with proteinase K added before incubation. DNA was purified from the resulting lysate with a QiaAmp MinElute Column (Qiagen, Hilden, Germany).

5.2.2.2 - Ampliseq panel sequencing

Library preparation was undertaken by the colleagues in the Stratified Medicine Core Laboratory using a custom Ampliseq panel (Thermo Fisher Scientific, Waltham, MA, USA) that included the *SDHA* region of interest. The protocol was adapted from a NEBNext Ultra II protocol for Illumina sequencing (New England Biolabs Inc., Ipswich, MA, USA). DNA samples were made up 10ng in 5µl and transferred to a 96 well plate with two primer pools (to avoid competition for hybridisation between adjacent primer pairs). Consequently, two wells were used per sample. Polymerase chain reactions (PCR) were performed by adding Q5 mastermix (New England Biolabs Inc., Ipswich, MA, USA) (Table 5.2) to each well and thermal cycling under the protocol described in Table 5.3. Following completion of PCR reactions, adaptor sequences were removed from amplicons by the addition of NEB USER Enzyme (New England Biolabs, Ipswich, MA, USA), which cleaves nucleic acids at uracil bases, and incubation with a thermal cycler. Wells corresponding to each sample for both primer pools were combined and transferred to wells of a MIDI plate containing 1.8X Agencourt AMpure XP magnetic beads (Beckman Coulter, Pasadena, CA, USA) to bind to amplicons. Two rounds of pull down and re-suspension were undertaken. To ligate specific barcode sequences to amplicons from specific samples, NEB End Repair reaction buffer then NEB End Repair enzyme mix (New England Biolabs Inc., Ipswich, MA, USA) was added to each well. 30µl NEB ligation master mix, 1µl of ligation enhancer and 2µl barcode sequence solution was added to each well before mixing and incubation. Further clean up using AMpure beads with ethanol washes were carried out. Quality of prepared libraries was measured by subjecting a 1/1000 dilution of each sample to quantitative PCR according to a KAPA protocol (Illumina Inc., San Diego, CA, USA). A 5µl aliquot of each sample was then transferred to an Illumina MiSeq instrument for sequencing.

Table 5.2 - PCR reaction components for Ampliseq panel

Reaction component	Volume (µl)
Q5 Master Mix	25
Primer Mix	10
DNA	5
Water	10
Total volume	50

Table 5.3 - PCR thermal cycling protocol for Ampliseq panel – 30 cycles

Step	Temperature (°C)	Duration (secs)
Initial denaturation	98	30
Denature	98	10
Anneal	60	30
Extend	65	120
Final extension	65	300
Hold	4	Hold

5.2.2.3 - Sanger sequencing

DNA extracted from tumours was also subject to Sanger sequencing for a *PALB2* variant identified in the corresponding blood DNA, performed by colleagues in the Department of Medical Genetics, University of Cambridge. PCR reactions for the region of interest were undertaken according to the advised protocol for AmpliTaq Gold DNA polymerase 50µl reaction (Thermo Fisher Scientific, Waltham, MA, USA). Primers are shown in Table 5.4 and reaction constituents are described in Table 5.5. Thermal cycling was performed on a Tetrad PTC-225 (MJ research, Waltham, MA, USA) according to the protocol described in Table 5.6. PCR products were subject to gel electrophoresis in 1% agarose gel (90v/40mins) and photographed under ultraviolet light to check for an observable band of predicted length. Following PCR, excess primers and deoxynucleotides were removed by adding a mixture of Exonuclease I (New England Biolabs, Ipswich, MA, USA) and Shrimp Alkaline Phosphatase (GE Healthcare, Chicago, IL, USA) to each PCR product well and incubating. Bidirectional Sanger sequencing of resulting products was performed with BigDye Terminator Version 3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) according to the reaction constituents described in Table 5.7 and thermal cycling protocol (with a Tetrad PTC-225) outlined in Table 5.8. To remove unincorporated dye, 40µl of 75% isopropanol was added to each well after the sequencing reaction. The plate containing the wells was then centrifuged and inverted onto absorbent paper to remove supernatant. It was left to air dry in dark conditions before adding 10µl of Hi-Di™ Formamide (Applied Biosystems, Foster City, CA, USA) to each well. The plate was then placed on an ABI 3131xl sequence analyser (Applied Biosystems, Foster City, CA, USA). Resulting chromatogram files were analysed with Sequencher 5.3 software (Gene Codes Corporation, Ann Arbor, MI, USA).

Table 5.4 - Primers used for amplifying region containing *PALB2* variant

Forward primer	CAACAGCAACACAAAACCACA
Reverse primer	AACTTTTGCTGAGGTCCAAGG

Table 5.5 - PCR reaction components for Sanger sequencing

Reaction component	Volume (µl)
AmpliTaq Gold DNA polymerase (5U/µl)	0.25
10µm Primer Mix	2
DNA	5
Water	33.75
10nM dNTP mix	1
25nM $MgCl_2$	3
10X PCR buffer	5
Total volume	50

Table 5.6 - PCR thermal cycling protocol for Sanger sequencing – 32 cycles

Step	Temperature (°C)	Duration
Initial denaturation	95	10 mins
Denature	95	15 secs
Anneal	59	30 secs
Extend	72	1 minute per kb
Final extension	72	5 mins
Hold	4	Indefinite

Table 5.7 - Sanger sequencing reaction components

Reaction component	Volume (µl)
BigDye Terminator Version 3.1	0.75
Primer solution (10pmol)	1
5x BigDye sequencing buffer	2
Water	4.25

Table 5.8 - Thermal cycling protocol for Sanger sequencing reaction – 20 cycles

Step	Temperature (°C)	Duration (secs)
Denature	96	10
Anneal	50	5
Extend	60	210

5.3 - Case reports

5.3.1 - Cases identified through sequencing studies

The following cases were identified through clinical practice of collaborators and/or sequencing studies undertaken in the Department of Medical Genetics, University of Cambridge.

FLCN*/*NF1

A 39 year old man presented with testicular seminoma and a routine abdominal scan four years later revealed a phaeochromocytoma. Following his seminoma diagnosis, he also developed a pneumothorax and went on to have six further occurrences. At age 55 years he complained of abdominal/ back pain and a computerised topography (CT) scan revealed bilateral renal masses that were demonstrated to be renal cell carcinomas (RCCs) following removal. Reinvestigation following further episodes of abdominal pain identified two gastrointestinal stromal tumours (GIST). At age 56 years, a CT lung scan (to investigate a pneumothorax) revealed a malignant peripheral nerve sheath tumour (MPNST). Skin examination revealed multiple skin neurofibromas, two café au lait patches and axillary freckling but no fibrofolliculomas. A clinical diagnosis of Neurofibromatosis type 1 was made and though this was considered to be the cause of his MPNST and possibly phaeochromocytoma and GIST, the history of renal cancers and recurrent pneumothorax were considered unrelated.

Next generation sequencing (NGS) of 94 CPGs was performed using the Illumina TruSight cancer panel.[169] A previously reported splice site variant in *FLCN* (ENST00000285071 c.1062+2T>G)[258,259] and a nonsense variant in *NF1* (ENST00000356175 c.1381C>T p.(Arg461*)) were detected and confirmed by Sanger sequencing. Deleterious *FLCN* variants cause Birt-Hogg-Dube syndrome (BHD), a rare condition where affected individuals are predisposed to RCC, pulmonary cysts, pneumothoraces and fibrofolliculomas. The patient's brother had also been diagnosed with bilateral chromophobe RCCs at age 45 years and was found to have facial fibrofolliculomas. Testing of a sister and her daughter demonstrated the presence of the *FLCN* variant but both were asymptomatic with normal renal scans. A paternal cousin with numerous fibrofolliculomas and a history of recurrent pneumothorax was confirmed to harbour the *FLCN* variant. The proband's deceased father had pancreatic adenocarcinoma but was not known to have features of BHD syndrome during life, although he was an obligate carrier of the *FLCN* variant and autopsy revealed bilateral renal oncocytomas. There was no known family history of Neurofibromatosis type 1.

Neurofibromatosis type 1 has a population frequency of 23/100.000[260] and might be expected to exist in combination with another inherited cancer syndrome relatively rarely, though phenotypic variability and use of clinical diagnostic criteria (rather than genetic testing) may underestimate this. It is associated with predisposition to a variety of neoplasms including phaeochromocytoma, GIST, carcinoid tumour, cutaneous/plexiform neurofibromas and MPNST. Thus, in this case associated with two pathogenic CPG variants, the occurrence of the MPNST, phaeochromocytoma, GIST and RCC can be explained but testicular seminoma has not been associated with variants in either gene.[149,261] This suggests that the seminoma might be a consequence of the combination of *FLCN* and *NF1* variants (seminoma has been linked to aberrations in the c-kit, RAS/MAPK and PI3K/AKTpathways[262,263] and the *NF1* and *FLCN* gene products regulate RAS/MAPK and mTOR/PI3K/Akt signalling respectively[264,265]) or be coincidental, testicular being the most common male solid tumour in the 15-34 age group.[266]

FLCN*/*TP53

A 32 year old man presented with dysphagia. There was a previous history of ulcerative colitis for which he had undergone a pan-protocolectomy at age 27 years and pathological examination of the colectomy specimen had revealed an incidental rectal adenocarcinoma. Endoscopy revealed a gastroesophageal junction adenocarcinoma and staging imaging demonstrated a 6cm left kidney tumour. Biopsy of the latter suggested a primary renal neoplasm, prompting nephrectomy. Histology of the resected kidney confirmed a chromophobe RCC. Examination of the skin showed facial fibrofolliculomas. There was no history of cancer in first degree relatives (both parents unaffected at age 60) but the maternal grandfather developed oesophageal squamous cell carcinoma at age 54. The

paternal grandmother and grandfather developed a brain tumour of uncertain histology and an oropharyngeal carcinoma at ages 50 and 49 years respectively.

Genetic investigations revealed two pertinent variants in *FLCN* (ENST00000285071 c.715C>T p.(Arg239Cys))[267] and *TP53* (ENST00000269305 c.526T>C p.(Cys176Arg)). The latter has been reported as a somatic mutational event on multiple occasions,[34,268] including in colorectal adenocarcinoma[34] but not previously in germline samples.[268] It is rare and does not appear in the Exome Aggregation Consortium (ExAC) dataset.[161] *In silico* tools predict a damaging or function altering effect.[269–271] No other family members were available for genetic testing.

Kidney tumours, typically with a hybrid chromophobe/oncocytic RCC histopathology, are a major feature of BHD syndrome. RCC has been reported in *TP53* pathogenic variant carriers though no firm association has been made.[56] It is noted that the median age at diagnosis of renal tumours in carriers of pathogenic *FLCN* variants (48 years)[259] is older than the age at onset of these tumours in this case, which might suggest a role for the *TP53* variant but rarity of BHD prevents accurate assessment of expected age of diagnosis. The relationship between colorectal cancer and BHD syndrome is controversial[258,272] but an increased risk of colorectal cancer has been reported with ulcerative colitis (though typically in those with disease for >10 years[273]) and also in carriers of pathogenic *TP53* variants.[274] To the author's knowledge, oesophageal cancers have not been reported in carriers of pathogenic *FLCN* variants but have occurred in Li-Fraumeni syndrome (LFS) families, though again the association with this condition is not clear.[56,190]

FLCN*/*MSH2

A 53 year old woman presented with rectal adenocarcinoma and had a history of spontaneous pneumothorax at age 46 years. Her father had developed colon cancer at 67 years and had several pneumothoraces (first at age 41 years). Immunohistochemistry performed on the proband's rectal tumour showed no abnormality but her father's colon cancer demonstrated loss of staining for MSH2 and MSH6 proteins. Constitutional genetic testing in the proband did not detect a pathogenic mismatch repair gene variant but a truncating *FLCN* variant (ENST00000285071 c.1285delC p.(His429Thrfs*39)) was identified. Three siblings had phenotypic similarities to the proband. A sister developed a pneumothorax at age 37 and had facial fibrofolliculomas. She also developed endometrial cancer at 52 years. Genetic testing demonstrated the familial *FLCN* variant and a truncating *MSH2* variant (ENST00000233146 c.892C>T p.(Gln298*)). The twin sister of this individual had pneumothoraces, RCC and colorectal polyps. She also carried both variants, as did a brother with facial fibrofolliculomas.

Colorectal and endometrial cancers are characteristic of Lynch syndrome (frequently caused by *MSH2*

variants) and the ages of diagnosis seen in this family are typical.[68] However, the proband did not carry the pathogenic *MSH2* variant detected in her siblings and may represent a phenocopy. Also, a role of the *FLCN* variant in the development of colorectal tumours in the family cannot be excluded.[258,272] Fibrofolliculomas, RCC and pneumothoraces are not associated with Lynch syndrome.[275]

XPA*/*MLH1

A male proband presented with a mucinous caecal cancer at age 65 years and a metachronous sigmoid colon cancer in his remaining large bowel at 67 years. He was one of eight siblings whose father had developed colon cancer at age 42 years, but there was no other family history of Lynch syndrome-related tumours. His parents were not knowingly consanguineous but were both from the same small community in India. The proband had been clinically diagnosed in early childhood with Xeroderma Pigmentosum (XP). His sister had a similar pattern of skin tumours but no internal malignancies. Neither of his parents had any reported skin abnormalities. On examination his sun-exposed skin showed considerable signs of ultraviolet damage (e.g. severe freckling and loss of pigment) but no other features of XP such as neurological or intellectual deficits. His skin tumours over the previous 20 years had included a squamous carcinoma in an actinic keratosis, several seborrheic keratoses, two keratoacanthomata/squamous carcinomas, junctional nevi, a squamous carcinoma and two lentigo malignae (premalignant melanoma). Immunohistochemistry demonstrated loss of MLH1 and PMS2 expression in both colon cancers. Constitutional genetic testing revealed *MLH1* ENST00000231790 c.306G>T p.(Glu102Asp) (classed as likely pathogenic[276]). Fibroblasts from a skin biopsy were tested for XP, which showed reduced levels of nucleotide excision repair. He therefore did not have mild XP variant (XP-V) as might be expected, but rather had mild variant XP-A, consistent with survival into his 60s. Constitutional genetic analysis revealed a homozygous *XPA* intron 4 splice variant (ENST00000375128 c.620+8A>G). Molecular analysis of his various tumours is summarised in Table 5.9.

Table 5.9 - Molecular analysis of tumours from *XPA*/*MLH1* case

Tumour	MLH1 IHC	PMS2 IHC	MSI assessment
Mucinous caecal adenocarcinoma	Loss	Loss	High
Sigmoid colon adenocarcinoma	Loss	Loss	High
Squamous carcinoma (#1)	Present	Present	Stable
Squamous carcinoma (#2)	Present	Present	Stable
Lentigo maligna	Present	Present	High
Actinic keratosis	Present	Present	High
Squamous carcinoma in actinic keratosis	Present	Present	High

MSI - Microsatellite instability. IHC - Immunohistochemistry

The prevalence of microsatellite instability (MSI) in skin tumours associated with XP is unknown. A contribution of the *MLH1* variant to the dermatological phenotype may be suggested by the MSI in some of the skin tumours but the presence of normal MLH1 and PMS2 expression goes against this. Skin tumours are associated with Lynch syndrome but these are characteristically sebaceous in origin, which were not observed in this case.

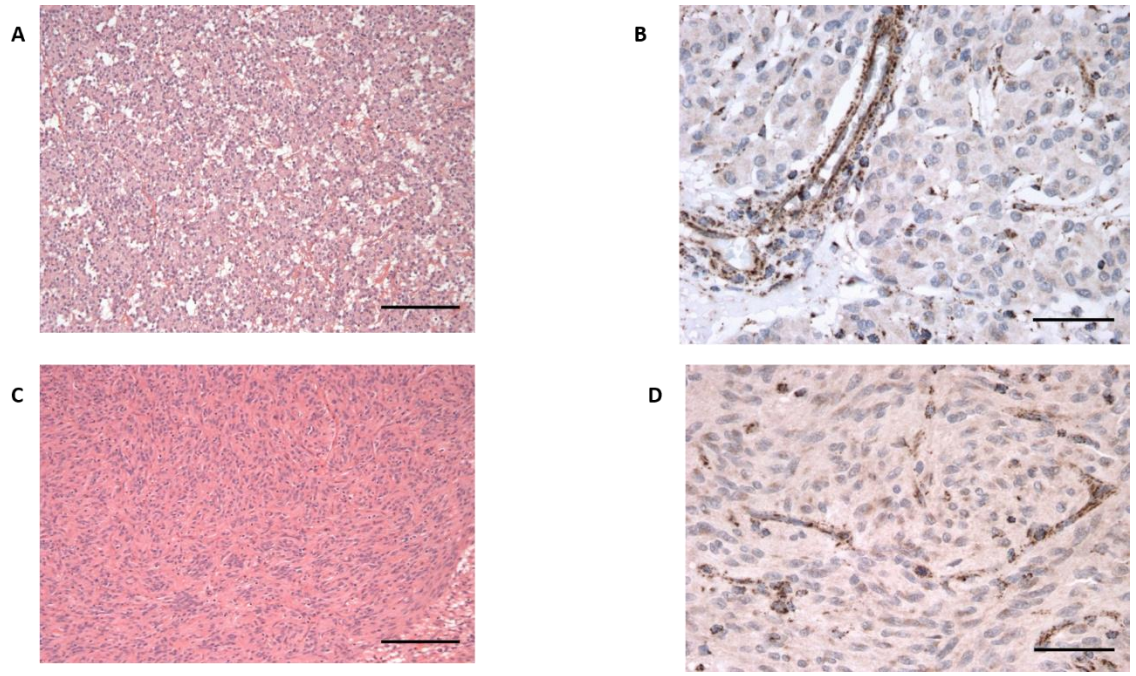
NF1*/*BRCA2

A female patient with Neurofibromatosis type 1, having one café au lait patch, numerous cutaneous neurofibromas, possible Lisch nodules and a MPNST, was diagnosed with ductal breast carcinoma at age 48 years and subsequently went on to develop a cutaneous melanoma at age 57 years. Constitutional genetic testing revealed both *NF1* ENST00000356175 c.6792C>G p.(Tyr2264*) and *BRCA2* ENST00000544455 c.5213_5216del p.(Thr1738Ilefs*2).[277] Pathogenic variants in both genes can be associated with breast cancer[278] but the risk is much higher for those affecting *BRCA2*. The breast cancer could be consistent with either syndrome and no tumour analysis was reported that could help determine which gene was more significant in its initiation.

SDHA*/*PALB2

A mother and son presented with GIST at age 66 and 34 respectively with the mother also developing breast cancer at age 70 years. Histology and immunohistochemistry of both GISTs showed a mixed epithelioid picture (expected in succinate dehydrogenase deficient GIST) and loss of SDHB staining, indicating inactivation of a component of the succinate dehydrogenase (SDH) complex (Figure 5.1). Constitutional genetic testing of *SDHX* genes showed a nonsense variant in *SDHA* (ENST00000264932 c.1532C>T (p.R512*)) in both individuals. These variants were confirmed by WGS undertaken on a research basis, which also identified *PALB2* ENST00000261584 c.3113G>A (p.Trp1038*) in both participants.

Figure 5.1 - Histology and SDHB immunohistochemistry on *SDHA*/*PALB2* diad. A and C show haematoxylin and eosin staining from son and mother respectively. B and D show loss of SDHB immunostaining in son and mother respectively.



Most GIST occurrences are sporadic and familial forms (known to have causes including constitutional variants in *KIT*, *PDGFRA*, *NF1* and *SDHX* genes) are rare.[279,280] This diad represents a further reported case of SDH deficient familial GIST. LOH analysis was performed on DNA from both tumours to confirm this and also investigate whether there was any evidence for the *PALB2* variant contributing to tumourigenesis. Loss of the *SDHA* wild type allele was confirmed with a panel-based sequencing assay where variant allele fraction (VAF) was 0.42 in the blood sample from the mother and 0.92 in her tumour sample. The son's samples showed VAF's of 0.57 in blood and 0.85 in tumour (Figure 5.2). Loss of the wild type *PALB2* allele, which may have indicated a contribution to increased penetrance of the *SDHA* variant and occurrence in two family members, was not observed (Figure 5.3). The *PALB2* variant is likely to have contributed to the breast cancer occurring in the mother but be incidental to the GIST occurrences. However, absence of LOH does not necessarily imply absent contribution (see below) and further tumour studies could potentially be revealing. Mutational signatures are derived from analysis of somatic single nucleotide variants and can provide insights into mutagenic processes leading to cancer in various tumour types.[281] One signature is associated with biallelic inactivation of *BRCA1* and *BRCA2* but has also has been demonstrated in breast[282] and pancreatic[283] cancers from individuals with constitutional *PALB2*

truncating variants. Analysis of signatures from these GISTs may show a similar signature but this appears unlikely given that the *PALB2* variants are not somatically biallelic and that GIST is not known to be associated with *PALB2* variants. Intriguingly however, succinate accumulation (which results from loss of succinate dehydrogenase function) has been reported to suppress DNA repair by homologous recombination.[284] This process is normally contributed to by functional *PALB2* protein product and loss of function variants in that gene lead to deficient repair. Feasibly, a concurrent *SDHA* variant could exacerbate that deficiency and promote tumourigenesis synergistically.

Figure 5.2 - Loss of *SDHA* wild type allele in familial GISTs

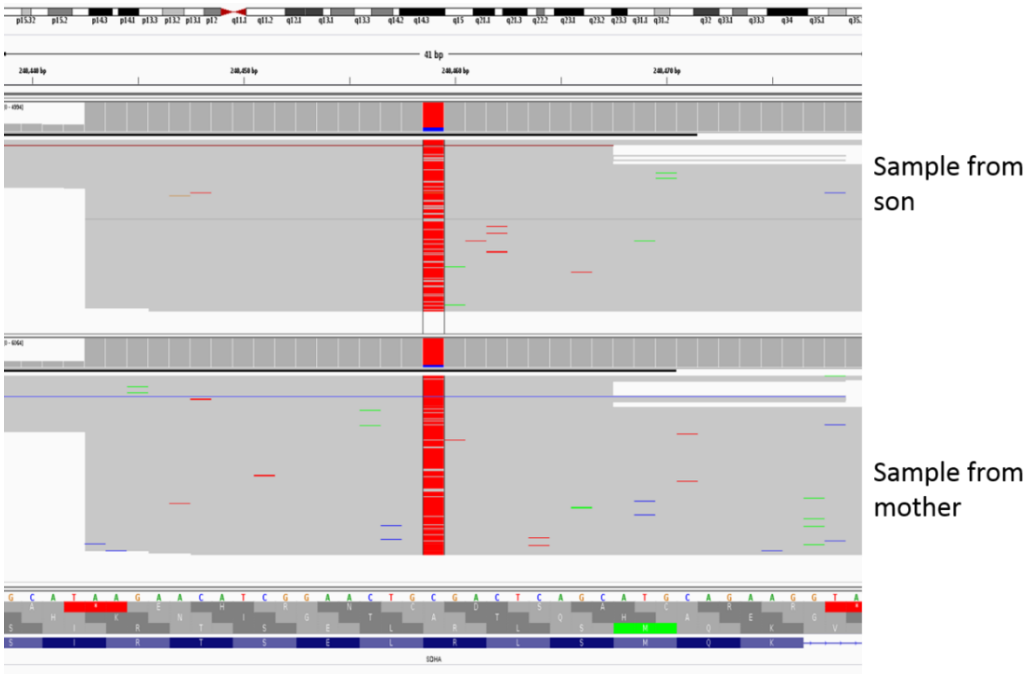
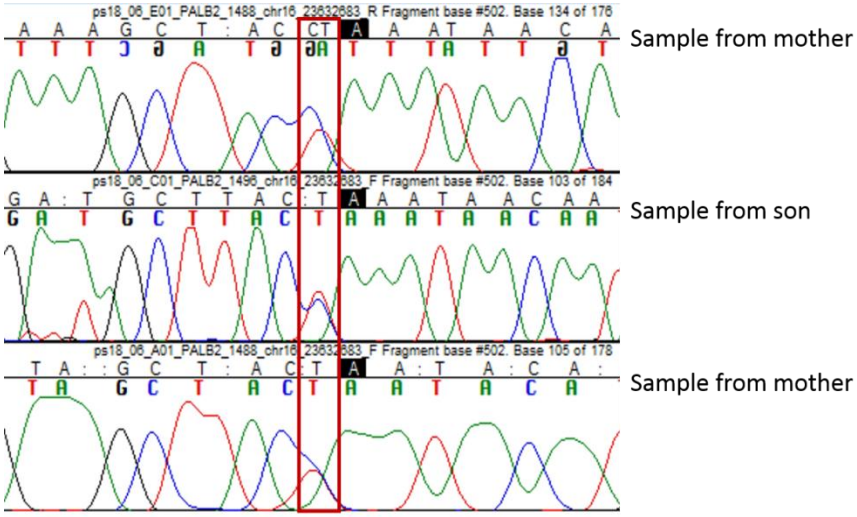


Figure 5.3 - Retention of *PALB2* wild type allele in familial GISTs



5.3.2 – Cases identified through whole genome sequencing-based comprehensive cancer predisposition gene analysis in multiple primary tumours series

Subsequently described cases were identified through analysis of WGS data from MPT individuals as described in Chapter 4.

PMS2/*BMPR1A*

An individual with colorectal adenocarcinoma at age 50 years and breast cancer at 57 years carried *PMS2* frameshift (ENST00000265849 c.741-742insTGAAG (p.Pro247_Ser248fs)) and *BMPR1A* nonsense (ENST00000372037 c.730C>T (p.Arg244*)) variants. Immunohistochemistry of the bowel tumour showed loss of PMS2 expression and microsatellite instability was demonstrated, leading to diagnostic sequencing of *PMS2.* There was no family history of neoplasia other than an ovarian cancer in a second degree relative after age 70 years. They had previously undergone surveillance colonoscopy for inflammatory bowel disease resulting in identification of a number of polyps but there was no evidence from histology reports that these were juvenile polyps. Given the results of the tumour studies and a polyp phenotype that is not highly characteristic of Juvenile Polyposis, the *PMS2* variant would appear likely to be causative of the colorectal adenocarcinoma but the role of either variant in development of the breast cancer is not clear.

MAX/*FH*

An MPT case with bilateral phaeochromocytoma at age 16 and 35 years with no reported family history of neoplasia was identified with *FH* (ENST00000366560 c.521C>G (p.Pro174Arg)) and *MAX* (ENST00000358664 c.1A>G (p.Met1Val)) variants.[285] The latter variant is predicted to abolish the *MAX* initiation codon and analysis of tumour tissue from an individual carrying it has previously demonstrated loss of the wild type allele and lack of full length MAX protein.[286] It is easier to attribute the diagnosed phaeochromocytoma to the truncating *MAX* variant but evidence for the role of *FH* in this tumour type is accumulating[209,210] and this variant may have contributed to tumourigenesis in either or both neoplasms.

FLCN/*CHEK2*

A further individual had the *CHEK2* ENST00000328354 c.1100delC (p.Thr367Metfs) variant (annotated in these data as ENST00000382580 c.1229delC (p.Thr410fs)) as well as a chromosome 17:10 translocation with a breakpoint within intron 9-10 of *FLCN*. Their phenotype included multiple cutaneous fibrofolliculomas and clear cell renal carcinoma at age 53 years. They had previously received a clinical diagnosis of BHD syndrome but sequencing of *FLCN* had not revealed any significant variants. The translocation appears to have been the causative factor for the fibrofolliculomas and renal cell carcinoma diagnosed in in this individual and the role of the *CHEK2* variant is unlikely to be significant in the development of the diagnosed tumours.

5.4 - Combination with cases from literature review

Combining the cases described above with those identified through literature review, 124 MINAS cases involving 29 CPGs were identified[200,239–252,254,256,257,287–319] (see Figure 5.4 and Table 5.10). 46 gene combination types were noted but only nine (*BRCA1*/*BRCA2*, *BRCA1/MLH1*, *BRCA1/CHEK2*, *BRCA2/CHEK2*, *FLCN/MSH2*, *APC/MSH2*, *ATM/BRCA1*, *BMPR1A/MSH2* and *APC/MLH1*) occurred in more than one family. This may reflect ascertainment bias (certain genes are commonly screened for simultaneously), common founder mutations present in specific populations and hereditary breast cancer, followed by colorectal cancer, being the most common indication for cancer genetic assessment.[173] Indeed, 13 individuals had a combination of two of the three Ashkenazi founder mutations in *BRCA1* and *BRCA2*.

Figure 5.4 - Combinations of pathogenic gene variants in MINAS cases from present report and literature review

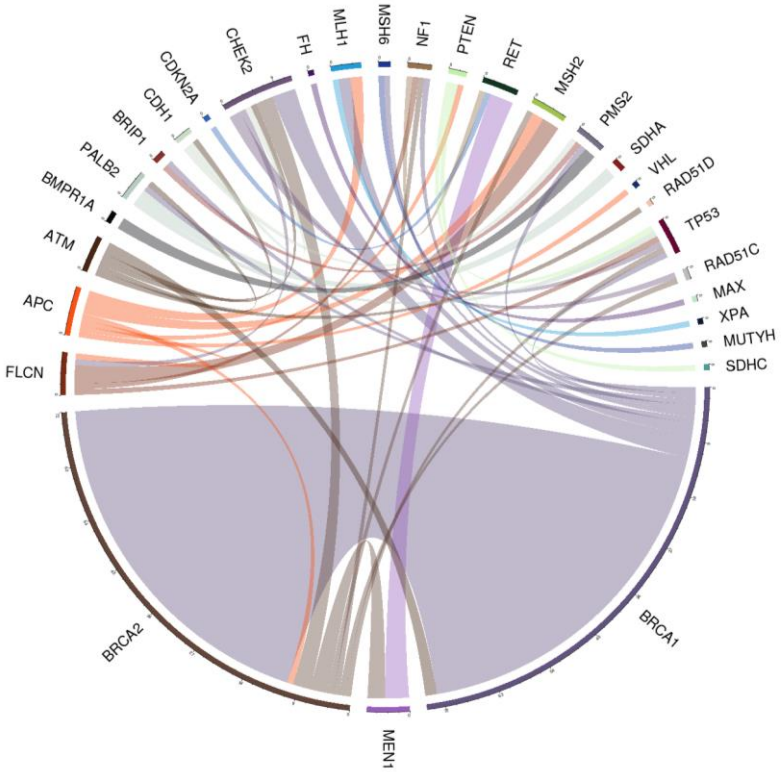


Table 5.10 - Multilocus Inherited Neoplasia Alleles Syndrome – details of published cases incorporating those in this report

Reference	Kindred within report	Sex	Gene 1	Gene 1 variant	Gene 2	Gene 2 variant	Clinical features with age in years at which noted (if known)
Goehringer et al. 2017	1	M	*APC*	ENST00000257430 c.3103dupC (p.Gln1035Profs*13)	*BRCA2*	ENST00000544455 c.516_516+1delGGinsT (p.Lys172Asnfs)	Intestinal polyposis, 35y†; Desmoid tumour (multiple), 36y†; Pancreatic adenocarcinoma, 54y‡
Kashiwada et al. 2012	1	F	*APC*	ENST00000257430 c.637C>T p.(Arg213*)	*FLCN*	ENST00000285071 c.1285dup p.(His429Profs*27)	Facial papules <28y‡; Colon carcinoma and multiple colon polyps 28y†; Recurrent pneumothoraces x4. Pulmonary cysts 28y (first one)‡
Lindor et al. 2012	1	M	*APC*	ENST00000257430 c.694C>T p.(Arg232*)	*MLH1*	Deletion exons 16-19	Rectal carcinoma and multiple colon polyps 14y*; Jejunal adenocarcinoma x6 28y x3, 34y, 44y, 52y (Loss of MLH1 and PMS2 on IHC)Δ; Duodenal adenocarcinoma 54y*; Congenital hypertrophy of retinal pigment epithelium 54y†; Squamous cell carcinoma. Multiple facialΔ; Pilomatricoma. Scalp 54y†; Sebaceous adenoma 54y (Loss of MLH1 and PMS2 on IHC)‡
Scheenstra et al. 2003	1	M	*APC*	ENST00000257430 c.3927_3931del p.(1309Aspfs*4)	*MLH1*	ENST00000231790 c.677G>A p.(Arg226Gln). Affects splicing.	Multiple colon polyps (100's) 10†; Tubular adenomas with dysplasia 10y (Loss of MLH1 on IHC)*
Soravia et al. 2005	1	M	*APC*	ENST00000257430 c.3471-3474delGAGA p.(Glu1157Aspfs*7)	*MSH2*	ENST00000233146 c.1192dupG p.(Ala398Glyfs*19)	Colon polyps x5. 4 adenomas 24y (1 dysplastic MSI high. Loss of MSH2 and MSH6 on IHC)†; Colon adenocarcinoma. Right colon 25y*; Gastric/duodenal adenoma x30 25y*; Desmoid tumour. Mesenteric 26y†
Uhrhammer and Bignon. 2008	1	M	*APC*	ENST00000257430 c.3183_3187delACAAA p.(Gln1062*)	*MSH2*	ENST00000233146 c.255_256del p.(Phe85Leufs*14)	Colon cancer 16y*
Kilmartin et al. 1996	1	M	*APC*	ENST00000257430 c.3340 C>T p.(Arg1114*)	*VHL*	Gene deletion (in offspring)	Retinal haemangioma x2 21y‡; Cerebellar haemangioblastoma 41y‡ Rectal carcinoma and multiple colonic polyps 41y†
Sokolenko et al. 2014	3	F	*ATM*	ENST00000278616 c.5932G>T (p.Glu1978*)	*BRCA1*	ENST00000357654 c.181T>G (p.Cys61Gly)	Breast cancer, 40y*
Sokolenko et al. 2014	4	F	*ATM*	ENST00000278616 c.5932G>T (p.Glu1978*)	*BRCA1*	ENST00000357654 c.181T>G (p.Cys61Gly)	Breast cancer, 42y*
Schrader et al. 2016	1	F	*ATM*	ENST00000278616 c.8793T>A (p.C2931*)	*CDH1*	ENST00000261769 c.1999delC (p.L667fs*12	Breast carcinoma 70-79y* (LOH ATM, CDH1 variant lost in tumour)

Sokolenko et al. 2014	5	F	*ATM*	ENST00000278616 c.5932G>T (p.Glu1978*)	*CHEK2*	del5395 (large deletion)	Breast cancer, 67y*
Crawford et al. 2017	1	F	*ATM*	ENST00000278616 c.901+1G >A	*PALB2*	ENST00000261584 c.2167_2168delAT (p.Met723Valfs)	Ovarian cancerΔ
Schrader et al. 2016	1	M	*ATM*	ENST00000278616 c.9139C>T (p.R3047*)	*RAD51D*	ENST00000345365 c.803G>A (p.W268*)	Non-small cell lung cancer 50-59yΔ (No LOH either variant)
This report	1	F	*BMPR1A*	ENST00000372037 c.730C>T (p.Arg244*)	*PMS2*	ENST00000265849 c.741-742insTGAAG (p.Pro247_S248fs)	Colorectal carcinoma, 50y*; Breast, 57yΔ
Silva-Smith et al. 2018	1	M	*BMPR1A*	ENST00000372037 c.25A > T (p.Arg9*)	*PMS2*	ENST00000265849 c.1882C>T (p.Arg628*)	Colorectal adenocarcinoma, 39y* (Loss of PMS2 immunostaining); Bladder transitional cell carcinoma, 39y‡
Augustyn et al. 2011	1	F	*BRCA1*	ENST00000357654 c.1961delA p.(Lys654Serfs*47)	*BRCA2*	ENST00000544455 c.1672delC p,(Ile558Leufs*15)	Ovarian serous carcinoma with papillary features. Bilateral 50y*
Augustyn et al. 2011	2	F	*BRCA1*	ENST00000357654 c.5266dupC p.(Gln1756Profs*74)	*BRCA2*	ENST00000544455 c.4829_4830del p.(Val1610Glyfs*4)	Breast cancer 40y (Triple negative histology)*
Bell et al. 2002	1	F	*BRCA1*	ENST00000357654 c.5266dupC p.(Gln1756Profs*74)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 33y (LOH BRCA2. No LOH BRCA1)‡; Breast cancer 44y (LOH BRCA2. No LOH BRCA1)‡; Breast cancer 47y (LOH BRCA1. No LOH BRCA2)†
Caldes 2002	1	F	*BRCA1*	ENST00000357654 c.5123C>A p.(Ala1708Glu)	*BRCA2*	ENST00000544455 c.6275_6276del p.(Leu2092Profs*7)	No features
Caldes 2002	2	M	*BRCA1*	ENST00000357654 c.5123C>A p.(Ala1708Glu)	*BRCA2*	ENST00000544455 c.6275_6276del p.(Leu2092Profs*7)	Prostate cancer 66y*
Caldes 2002	1	F	*BRCA1*	ENST00000357654 c.5123C>A p.(Ala1708Glu)	*BRCA2*	ENST00000544455 c.6275_6276del p.(Leu2092Profs*7)	Breast cancer 70y*
Caldes 2002	1	F	*BRCA1*	ENST00000357654 c.5123C>A p.(Ala1708Glu)	*BRCA2*	ENST00000544455 c.6275_6276del p.(Leu2092Profs*7)	Breast cancer 66y*
Caldes 2002	1	F	*BRCA1*	ENST00000357654 c.5123C>A p.(Ala1708Glu)	*BRCA2*	ENST00000544455 c.6275_6276del p.(Leu2092Profs*7)	Breast cancer 28y (No LOH BRCA1 or BRCA2)*

Caldes 2002	1	F	*BRCA1*	ENST00000357654 c.5123C>A p.(Ala1708Glu)	*BRCA2*	ENST00000544455 c.6275_6276del p.(Leu2092Profs*7)	No features
Caldes 2002	1	F	*BRCA1*	ENST00000357654 c.5123C>A p.(Ala1708Glu)	*BRCA2*	ENST00000544455 c.6275_6276del p.(Leu2092Profs*7)	No features
Caldes 2002	1	F	*BRCA1*	ENST00000357654 c.5123C>A p.(Ala1708Glu)	*BRCA2*	ENST00000544455 c.6275_6276del p.(Leu2092Profs*7)	No features
Choi et al. 2006	1	F	*BRCA1*	ENST00000357654 c.1504_1508delTTAAA p.(Leu502Alafs*2)	*BRCA2*	ENST00000544455 c.2798_2799delCA p.(Thr933Argfs*2)	Breast infiltrating duct carcinoma 26y*
Choi et al. 2006	2	F	*BRCA1*	ENST00000357654 c.4981G>T p.(Glu1661*)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast infiltrating duct carcinoma 33y*
Heidemann et al. 2012	1	F	*BRCA1*	ENST00000357654 c.5266dup p.(Gln1756Profs*74)	*BRCA2*	ENST00000544455 c.5645C>G p.(Ser1882*)	Breast cancer 37y*; Breast cancer 39y*; Ovarian cancer 63y*
Heidemann et al. 2012	2	M	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5718_5719delCT (p.Ser1907Leufs*4)	No features
Heidemann et al. 2012	2	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5718_5719delCT (p.Ser1907Leufs*4)	Breast cancer 32y*
Heidemann et al. 2012	3	F	*BRCA1*	ENST00000357654 c.962G>A p.(Trp321*)	*BRCA2*	ENST00000544455 c. 2231C>G p.(Ser744*)	Breast cancer 31y*; Breast cancer (contralateral) 35y*
Heidemann et al. 2012	4	F	*BRCA1*	ENST00000357654 c.3910delG p.(Glu1304Lysfs*3)	*BRCA2*	ENST00000544455 c.2830A>T p.(Lys944*)	Breast cancer 39y*
Heidemann et al. 2012	5	F	*BRCA1*	ENST00000357654 c.5277+1delG	*BRCA2*	ENST00000544455 c.658_659delGT p.(Val220Ilefs*4)	Colorectal cancer. Caecal 58yΔ; Ovarian cancer 61y*
Heidemann et al. 2012	5	M	*BRCA1*	ENST00000357654 c.5277+1delG	*BRCA2*	ENST00000544455 c.658_659delGT p.(Val220Ilefs*4)	No features
Heidemann et al. 2012	6	F	*BRCA1*	ENST00000357654 c.3700_3704delGTAAA p.(Val1234Glnfs*8)	*BRCA2*	ENST00000544455 c.1813_1814insA p.(Ile605Asnfs*11)	Cervical cancer 26yΔ; Breast cancer 40y*

Leegte et al. 2005	1	F	*BRCA1*	ENST00000357654 c.2685_2686delAA p.(Pro897fs*5)	*BRCA2*	ENST00000544455 c.3487delG p.(Asp1163Ilefs*5)	Ovarian papillary serous cystadenocarcinoma 40y (LOH BRCA2)*; Breast infiltrative ductal carcinoma 45y (LOH BRCA1)*
Leegte et al. 2005	2	F	*BRCA1*	ENST00000357654 c.2685_2686delAA p.(Pro897fs*5)	*BRCA2*	ENST00000544455 c.4449delA p.(Asp1484Thrfs*2)	Breast cancer. Ductal 28y*
Leegte et al. 2005	3	F	*BRCA1*	ENST00000357654 c.66_67delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	No features
Leegte et al. 2005	4	F	*BRCA1*	ENST00000357654 c.5263_5264insC p.(Ser1756Profs*74)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast invasive lobular carcinoma 51y*
Liede et al. 1998	1	F	*BRCA1*	ENST00000357654 c.2389G>T p.(Glu797*)	*BRCA2*	ENST00000544455 c.3067_3068insA p.(Asn1023Lysfs*3)	Breast adenocarcinoma 35y*
Loubser et al. 2012	1	M	*BRCA1*	ENST00000357654 c.2641G>T p.(Glu881*)	*BRCA2*	ENST00000544455 c.7934delG p.(Arg2645Asnfs*3)	No features 49y
Loubser et al. 2012	1	F	*BRCA1*	ENST00000357654 c.2641G>T p.(Glu881*)	*BRCA2*	ENST00000544455 c.7934delG p.(Arg2645Asnfs*3)	Breast ductal carcinoma 42y
Moslehi et al. 2000	1	F	*BRCA1*	ENST00000357654 c.66_67delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	No features 36y
Musolino et al. 2005	1	F	*BRCA1*	ENST00000357654 c.4285_4286insG p.(Tyr1429*)	*BRCA2*	ENST00000544455 c.7738C>T p.(Gln2580*)	Breast infiltrating duct carcinoma 37y (Triple negative histology)*
Noh et al. 2011	1	F	*BRCA1*	ENST00000357654 c.3746_3747insA p.(Glu1250Argfs*5)	*BRCA2*	ENST00000544455 c.6952_6953del p.(Arg2318Lysfs*21)	Breast infiltrating duct carcinoma 26y*
Noh et al. 2011	2	F	*BRCA1*	ENST00000357654 c.390C>A p.(Tyr130*)	*BRCA2*	ENST00000544455 c.3018delA p.(Gly1007Valfs*36)	Breast infiltrating duct carcinoma 45y*
Noh et al. 2011	3	F	*BRCA1*	ENST00000357654 c.5030_5033delCTAA p.(Thr1677Ilefs*2)	*BRCA2*	ENST00000544455 c.1399A>T p.(Lys467*)	Breast infiltrating duct carcinoma 35y*

Pilato et al. 2010	1	F	*BRCA1*	ENST00000357654 c.5266dup p.(Gln1756Profs*10)	*BRCA2*	ENST00000544455 c.5796_5797delTA p.(His1932Glnfs*12)	Breast intraductal carcinoma 38y (Triple negative histology)*; Ovarian papillary adenocarcinoma. Bilateral 42y*
Zuradelli et al. 2010	2	F	*BRCA1*	ENST00000357654 c.3916_3917delTT p.(Leu1306Aspfs*23)	*BRCA2*	ENST00000544455 c.5380delG p.(Val1794*)	Breast ductal cancer. Medullary type 30y (ERPR-ve)*; Ovarian serous papillary carcinoma 36y*
Zuradelli et al. 2010	3	F	*BRCA1*	ENST00000357654 c.1687C>T p.(Gln563*)	*BRCA2*	ENST00000544455 c.6469C>T p.(Gln2157*)	Breast infiltrating duct carcinoma 2x foci 46y (1 lymph node ERPR -ve. 1 lymph node ERPR+ve)*
Zuradelli et al. 2010	4	F	*BRCA1*	ENST00000357654 c.2405_2406delTG p.(Val802Glufs*7)	*BRCA2*	ENST00000544455 c.4285C>T p.(Gln1429*)	Breast ductal carcinoma 52y (Triple negative histology)*; Ovarian serous adenocarcinoma. Bilateral 52y*
Friedman et al. 1998	1	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 38y*
Friedman et al. 1998	2	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Ovarian cancer 57y*
Friedman et al. 1998	3	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	No features
Friedman et al. 1998	4	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 45y*
Ramus et al. 1997	1	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 48y*; Ovarian cancer 50y*
Leegte et al. 2005/Frank et al. 2002	Unknown	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 39y*
Leegte et al. 2005/Frank et al. 2002	Unknown	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 41y*
Leegte et al. 2005/Frank et al. 2002	Unknown	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer. Bilateral 34y*
Leegte et al. 2005/Frank et al. 2002	Unknown	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 55y*; Breast cancer (contralateral) 56y*

Leegte et al. 2005/Frank et al. 2002	Unknown	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	No features
Leegte et al. 2005/Frank et al. 2002	Unknown	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 40y*
Leegte et al. 2005/Frank et al. 2002	Unknown	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer 33y*; Breast cancer (contralateral) 49y*
Leegte et al. 2005/Frank et al. 2002	Unknown	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	No features 61y
Randall et al. 1998	1	F	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer. Multifocal lobular carcinoma 30y (LOH BRCA1)*; Ovarian cancer 41y (LOH BRCA1 and BRCA2)*
Meynard et al. 2017	1	F	*BRCA1*	ENST00000357654 c.1016dupA (p.V340Glyfs*6)	*BRCA2*	ENST00000544455 c.6814delA (p.Arg2272Glufs)	Bilateral breast carcinoma, 46y*
Nomizu et al. 2015	1	F	*BRCA1*	ENST00000357654 c.188T>A (p.Leu63*)	*BRCA2*	ENST00000544455 c.5576delTTAA (p.Ile1861fs)	Breast carcinoma, 55y* (Negative immunostaining for BRCA1 and BRCA2)
Nomizu et al. 2015	1	F	*BRCA1*	ENST00000357654 c.188T>A (p.Leu63*)	*BRCA2*	ENST00000544455 c.5576delTTAA (p.Ile1861fs)	Breast cancer, 41y*; Endometrial cancer, 46yΔ
Vietri et al. 2013	1	F	*BRCA1*	ENST00000357654 c.547+2T>A	*BRCA2*	ENST00000544455 c.2830A>T (p.Lys944*)	Bilateral breast cancer, 43y*
Vietri et al. 2013	1	M	*BRCA1*	ENST00000357654 c.547+2T>A	*BRCA2*	ENST00000544455 c.2830A>T (p.Lys944*)	No tumours, 72y
Vietri et al. 2013	1	F	*BRCA1*	ENST00000357654 c.547+2T>A	*BRCA2*	ENST00000544455 c.2830A>T (p.Lys944*)	Breast cancer,39y*
Schrader et al. 2016	1	M	*BRCA1*	ENST00000357654 c.68_69delAG p.(Glu23Valfs*17)	*BRIP1*	ENST00000259008 c.1315C>T (p.R439*)	Testicular seminoma 20-29yΔ (No LOH either variant)
Sokolenko et al. 2014	1	F	*BRCA1*	ENST00000357654 c.5266dupC p.(Gln1756Profs*74)	*CHEK2*	del5395 (large deletion)	Breast cancer, 52y (LOH CHEK2, No LOH BRCA1)*
Sokolenko et al. 2014	2	F	*BRCA1*	ENST00000357654 c.3247_3251delATGCT (p.Met1083fs)	*CHEK2*	del5395 (large deletion)	Breast cancer, 42y*

Sokolenko et al. 2014	6	F	*BRCA1*	ENST00000357654 c.181T>G (p.Cys61Gly)	*CHEK2*	ENST00000382580 c.1229delC (p.Thr410fs)	Breast cancer, 58y*
Sokolenko et al. 2014	7	F	*BRCA1*	ENST00000357654 c.5266dupC p.(Gln1756Profs*74)	*CHEK2*	ENST00000382580 c.396+1G>T	Breast cancer, 54y*
Pedroni et al. 2013	1	F	*BRCA1*	ENST00000357654 c. 181T>G p.(Cys61Gly)	*MLH1*	ENST00000231790 c.1489dupC p.(Arg497Profs*6)	Breast cancer 35y (Loss of MLH1 on IHC. LOH MLH1 and BRCA1)†; Endometrial carcinoma (Loss of MLH1 on IHC. LOH MLH1)‡; Ovarian carcinoma 39y (Loss of MLH1 on IHC. LOH MLH1)*; Renal clear cell carcinoma 39yΔ; Breast cancer (contralateral) 46y (Loss of MLH1 on IHC. LOH and BRCA1)‡
Kast et al. 2012	1	F	*BRCA1*	ENST00000357654 c.213-12A>G, p.(Arg71Serfs*21). Cryptic splice site	*MSH6*	ENST00000234420 c.515dup p.(Leu173Thrfs*9)	Endometrial endometrioid adenocarcinoma 46y (Loss of MSH6 on IHC)‡
Campos et al. 2013	1	F	*BRCA1*	ENST00000357654 c.4107_4110dup p.(Gly1371Ilefs*4)	*NF1*	ENST00000356175 c.4120C>T p.(Gln1374*)	Café au lait patches multiple cutaneous neurofibromas and Axillary/inguinal freckling in childhood‡; Breast infiltrating duct carcinoma 35y‡
Pern et al. 2012	1	F	*BRCA1*	ENST00000357654 c.927delA p.(Lys309Asnfs*5)	*PALB2*	ENST00000261584 c.756dup p.(Leu253Serfs*4)	Uterine myomas <65yΔ; Meningioma <65yΔ; Breast invasive ductal carcinoma. Multifocal 65y (Triple negative histology)*
Eliade et al. 2017	1	F	*BRCA1*	ENST00000357654 c.1480C>T (p.Gln494*)	*PMS2*	ENST00000265849 c.251-2A>T	Breast cancer, 65y†; Ovarian cancer, 72y†
Bell et al. 2014	1	F	*BRCA1*	ENST00000357654 c.81-?_134 + ? del (p.Cys27*) (exon 3 deletion)	*TP53*	ENST00000269305 c.375 + 2T > C	Breast carcinoma, 20y*
Smith et al. 2008	1	F	*BRCA1*	ENST00000357654 c.3331_3334delCAAG p.(Gln1111Asnfs*5)	*BRCA2*	ENST00000544455 c.631+2T>G	Breast cancer 34y*; Colorectal carcinoma. Transverse. (No loss of MMR proteins on IHC. No microsatellite instability) 35yΔ; Breast cancer 53y*
Smith et al. 2008	1	F	*BRCA1*	ENST00000357654 c.3331_3334delCAAG p.(Gln1111Asnfs*5)	*BRCA2*	ENST00000544455 c.631+2T>G	No features 65y
Tesoriero et al. 1999	1	F	*BRCA1*	ENST00000357654 c.3769_3770delGA p.(Glu1257Glyfs*9)	*BRCA2*	ENST00000544455 c.5946delT p.(Ser1982Argfs*22)	Breast cancer <40y (LOH BRCA2)*
Zuradelli et al. 2010	1	F	*BRCA1*	ENST00000357654 c.835delC p.(His279Metfs*19)	*BRCA2*	ENST00000544455 c.8195T>G p.(Leu2732*)	Breast carcinoma. Metaplastic 43y (Triple negative histology)*

Borg et al. 2000	1	F	*BRCA1*	ENST00000357654 c.3047_3048insTGAGA p.(Asn1018Metfs*8)	*MLH1*	ENST00000231790 c.131C>T p.(Ser44Phe). Additional non-pathogenic variant c.1321G>A p.(Ala441Thr)	Breast invasive ductal carcinoma 35y (MSI low. ERPR -ve)†
Eliade et al. 2017	3	F	*BRCA2*	ENST00000544455 c.6952C>T (p.Arg2318*)	*CHEK2*	ENST00000328354 c.1427C>T (p.Thr476Met). 7 ClinVar reports LP (more recent), 7 reports VUS	No tumours
Francies et al. 2015	1	F	*BRCA2*	ENST00000544455 c.5213_5216delCTTA (p.Thr1738Ilefs)	*CHEK2*	ENST00000382580 c.1229delC (p.Thr410fs)	Breast cancer, <50y*
Schrader et al. 2016	1	F	*BRCA2*	ENST00000544455 c.3846_3847delTG (p.V1283fs*2)	*CHEK2*	ENST00000328354 c.793-1G>A	Breast carcinoma 50-59y* (No LOH either variant)
Ghataorhe et al. 2007	1	F	*BRCA2*	ENST00000544455 c.2808_2811delACAA p.(Ala938Profs*21)	*MEN1*	ENST00000312049 c. 1159+1delGT	Abnormal secretory parathyroid gland 34y‡; Pancreatic mass. Unknown histology. Non-functional 35y*
Ghataorhe et al. 2007	1	F	*BRCA2*	ENST00000544455 c.2808_2811delACAA p.(Ala938Profs*21)	*MEN1*	ENST00000312049 c. 1159+1delGT	Cushing syndrome (implied pituitary origin) 10y‡; Hypercalcaemia (implied hyperparathyroidism) 31y‡
Ghataorhe et al. 2007	1	M	*BRCA2*	ENST00000544455 c.2808_2811delACAA p.(Ala938Profs*21)	*MEN1*	ENST00000312049 c. 1159+1delGT	Parathyroid hyperplasia 56y‡; Breast cancer 60y†
Thiffault et al. 2004	1	F	*BRCA2*	ENST00000544455 c.314T>G p.(Leu105*)	*MSH2*	ENST00000233146 c.1277_1386del (Exon 8 deletion)	Lobular and ductal carcinoma in situ 32y (ERPR +ve)†; Endometrioid adenocarcinoma 40y (No MMR deficiency on IHC. MSI low)Δ; Colon villotubular adenoma. 40 (Loss of MSH2 on IHC. MSI high)‡
This report	1	F	*BRCA2*	ENST00000544455 c.5213_5216del p.(Thr1738Ilefs*2)	*NF1*	ENST00000356175 c.6792C>G p.(Tyr2264*)	Breast ductal carcinoma 48y*; Cutaneous melanoma 57y†; Multiple cutaneous neurofibromas‡; Malignant peripheral nerve sheath tumour‡; Café au lait patch‡; Possible Lisch nodules‡.
Ahlborn et al. 2015	1	F	*BRCA2*	ENST00000544455 c.9648G>A (p.Leu3216=) (abnormal splicing)	*RAD51C*	ENST00000337432 c.773G>A (p.Arg258His)	No tumours, 38y
Monnerat et al. 2007	1	F	*BRCA2*	ENST00000544455 c.4889C>G p.(Ser1630*)	*TP53*	ENST00000269305 c.329G>T p.(Arg110Leu)	Cutaneous malignant melanoma 65y‡; Breast cancer 69y*; Ovarian cancer 69y*; Colon cancer 74y‡

Schrader et al. 2016	1	F	*BRIP1*	ENST00000259008 c.2392C>T (p.R798*)	*PMS2*	ENST00000265849 c.137G>T (p.Ser46Ile)	Breast carcinoma 80-89yΔ (No LOH PMS2, BRIP1 variant lost in tumour)
Schrader et al. 2016	1	F	*CDH1*	ENST00000261769 c.1090_1105dupACAGTCACTGACACCA (p.D370fs*3)	*CHEK2*	ENST00000382580 c.1229delC (p.Thr410fs)	Oesophageal adenocarcinoma 50-59yΔ (No LOH CHEK2, CDH1 variant lost in tumour)
Njoroge et al. 2017	1	F	*CDH1*	ENST00000261769 c.2287G>T (p.Glu763*)	*PMS2*	ENST00000265849 c.2445+1G>T	Breast carcinoma (lobular), 51y† (loss of e-cadherin immunostaining); Thyroid papillary carcinoma, 52yΔ
This report	1	M	*CHEK2*	ENST00000382580 c.1229delC (p.Thr410fs)	*FLCN (SV)*	ENST00000285071 17:10 translocation with a breakpoint within intron 9-10	Fibrofolliculoma (multiple), 18y‡; Renal cell carcinoma, 53y‡
Crawford et al. 2017	2	F	*CHEK2*	ENST00000382580 c.1229delC (p.Thr410fs)	*RAD51C*	ENST00000337432 c.397C>T (p.Gln133*)	Ovarian cancer‡
Schrader et al. 2016	1	F	*CHEK2*	ENST00000328354 c.470T>C (p.Iel157Thr). 12 ClinVar reports P/LP, 3 reports VUS	*TP53*	ENST00000269305 c.505_506delAT (p.M169fs*11)	Soft tissue sarcoma 50-59y‡ (CHEK2 variant lost in tumour)
This report	1	F	*FH*	ENST00000366560 c.521C>G (p.Pro174Arg)	*MAX*	ENST00000358664 c.1A>G (p.Met1Val)	Phaeochromocytoma, 16y*; Phaeochromocytoma, 35y*
This report	1	F	*FLCN*	ENST00000285071 c.1285delC p.(His429Thrfs*39)	*MSH2*	ENST00000233146 c.892C>T p.(Gln298*)	Pneumothorax 37y†; Endometrial cancer 52y‡.
This report	1	F	*FLCN*	ENST00000285071 c.1285delC p.(His429Thrfs*39)	*MSH2*	ENST00000233146 c.892C>T p.(Gln298*)	Renal cell carcinoma†; Colorectal polyps‡; Multiple pneumothoraces†.
This report	1	M	*FLCN*	ENST00000285071 c.1285delC p.(His429Thrfs*39)	*MSH2*	ENST00000233146 c.892C>T p.(Gln298*)	Facial fibrofolliculomas†
This report	1	M	*FLCN*	ENST00000285071 c.1062+2T>G	*NF1*	ENST00000356175 c.1381C>T p.(Arg461*)	Testicular seminoma 39yΔ; Renal cell carcinoma. Chromophobe 55y†; Phaeochromocytoma 43y‡; Gastrointestinal stromal tumour x2 55y‡; Malignant peripheral nerve sheath tumour 56y‡; Multiple cutaneous neurofibromas‡; Cafe au lait patches‡; Recurrent pneumothoraces.
This report	1	M	*FLCN*	ENST00000285071 c.715C>T p.(Arg239Cys)	*TP53*	ENST00000269305 c.526T>C p.(Cys176Arg)	Rectal carcinoma 27yΔ; Gastroesophageal adenocarcinoma 32yΔ; Renal cell carcinoma. Chromophobe 32y†; Facial fibrofolliculomas†.

This report	1	M	*MLH1*	ENST00000231790 c.306G>T p.(Glu102Asp)	*XPA*	ENST00000375128 c.620+8A>G	Caecal cancer. Mucinous 65y†; Sigmoid cancer 67y†; Previous skin tumours including squamous carcinoma in an actinic keratosis, multiple seborrhoeic keratoses, keratoacanthomata/squamous carcinomas x2, junctional naevi, squamous carcinoma and lentigo malignae x2‡
Puijenbroek et al. 2007	1	F	*MSH6*	ENST00000234420 c.1784delT p.(Leu595fs*15)	*MUTYH (compound heterozygote)*	ENST00000450313 c.536A>G p.(Tyr179Cys) and c.1187G>A p.(Gly396Asp)	Colon adenomas x5 48y (All MSI stable. Retained MSH6 expression)‡
Ercolino et al. 2014	1	M	*NF1*	ENST00000356175 c.1185+1G>A	*RET*	ENST00000355710 c.2410G>A p.(Val804Met)	Macrocephaly, café au lait patches and axillary freckling 57y†; Kyphoscoliosis 57y†; Multiple cutaneous neurofibromas 57y†; Thyroid C-cell hyperplasia 57y‡; Parathyroid hyperplasia 57y‡
This report	1	M	*PALB2*	ENST00000261584 c.3113G>A (p.Trp1038*)	*SDHA*	ENST00000264932 c.91C>T (p.R31*)	GIST (gastric), 66y‡ (Loss of SDHB immunostaining and LOH SDHA); Breast DCIS, 70y†
This report	1	F	*PALB2*	ENST00000261584 c.3113G>A (p.Trp1038*)	*SDHA*	ENST00000264932 c.91C>T (p.R31*)	GIST (gastric), 34y‡ (Loss of SDHB immunostaining and LOH SDHA)
Eliade et al. 2017	2	F	*PALB2*	ENST00000261584 c.1135A>T (p.Lys379*)	*TP53*	ENST00000269305 c.743G>A (p.Arg248Gln) and c.473C>T (p.Arg158His)	Ovarian cancer, 41yΔ; Breast cancer, 61y*; Pancreatic cancer, 63y†
Valle et al. 2004	1	F	*PTEN*	ENST00000371953 c.634+5G>A	*APC*	ENST00000257430 c.541insA p.(Gln181Thrfs*12)	Multiple colonic polyps 10y‡; Subcutaneous nodules‡; Multinodular goitre 26y†; Papillary thyroid cancer, multiple nodular hyperplasia and follicular adenomas 26y†; Diffuse lymphocytic chronic thyroiditis†; Ovarian Morgani hydatid 15yΔ; Cerebellar dysplastic gangliocytoma 26y†; Palmar keratosis 26y†; Head fibroma 26y†; Lipomas 26y†; Melanocytic naevi x2 28y†; Facial papules 28y†; Oral papillomatosis 28y†
Zbuk et al. 2007	1	F	*PTEN*	ENST00000371953 c.47dup p.(Tyr16*)	*SDHC*	ENST00000367975 c.397C>T p.(Arg133*)	Macrocephaly†; Papillomatous papules†; Paraganglioma. Left common carotid 18y‡; Fibrocystic breast disease 20's†; Papillary thyroid cancer 37y†; Paraganglioma. Right carotid body 39y‡; Uterine leiomyomas 30's†

Plon et al. 2008	1	F	*PTEN*	ENST00000371953 c.334C>G p.(Leu112Val). Cryptic splice site	*TP53*	ENST00000269305 c.844C>T p.(Arg282Trp)	Neuroblastoma 0y‡; Lipoma. Abdominal wall 0y†; Haemangiomas 1y†; Macrocephaly†; Ovarian granulosa cell tumour 1y (No somatic PTEN or TP53 variants. LOH PTEN. No LOH TP53)Δ; Xanthoastrocytoma. Temporal lobe 3y (No somatic PTEN or TP53 variants. No LOH PTEN or TP53)Δ; Pelvic liposarcoma 4y (No somatic PTEN or TP53 variants. LOH PTEN. No LOH TP53)Δ
Foppiani et al. 2008	1	M	*RET*	ENST00000355710 c.2410G>A p.(Val804Met)	*CDKN2A*	ENST00000304494 c.142C>A p.(Pro48Thr). ClinVar single submitter	Cutaneous malignant melanoma <55y‡; Parathyroid chief cell adenoma 55y†; Thyroid sclerotic papillary carcinoma 55y†; Thyroid C cell hyperplasia 55y†
Mastroianno et al. 2011	1	M	*RET*	ENST00000355710 c.1997A>T p.(Lys666Met)	*MEN1*	ENST00000312049 c.893+1G>T	Pituitary tumour 38y‡; Primary hyperparathyroidism 45y*; Papillary thyroid cancer 46yΔ; Medullary thyroid cancer 46y†; Gastric carcinoid tumour 47y‡; Gastrinoma‡
Mastroianno et al. 2011	1	M	*RET*	ENST00000355710 c.1997A>T p.(Lys666Met)	*MEN1*	ENST00000312049 c.893+1G>T	Primary hyperparathyroidism 40y*; Cushing syndrome (implied pituitary origin) 40y‡; Carcinoid tumour 40y‡; Lipoma 40y‡; Angiofibroma 40y‡; Papillary thyroid cancerΔ; Medullary thyroid cancer 40y†; Gastrinoma 41y‡
Mastroianno et al. 2011	1	M	*RET*	ENST00000355710 c.1997A>T p.(Lys666Met)	*MEN1*	ENST00000312049 c.893+1G>T	No features 6y
Mastroianno et al. 2011	1	F	*RET*	ENST00000355710 c.1997A>T p.(Lys666Met)	*MEN1*	ENST00000312049 c.893+1G>T	Primary hyperparathyroidism 13y*; Pituitary tumour 15y*

†Tumour type associated with pathogenic variants in gene 1
‡Tumour type associated with pathogenic variants in gene 2
*Tumour type associated with gene 1 and gene 2
Δ Tumour type associated with pathogenic variants in neither gene 1 or gene 2

LOH - Loss of heterozygosity i.e. loss of normal allele for quoted gene in tumour, IHC - Immunohistochemistry, ER - Oestrogen receptor, PR - Progesterone receptor, VUS – Variant of uncertain significance, MMR – Mismatch repair

5.5 - Discussion

5.5.1 - Delineating the relative significance of variants through molecular investigation

In theory, insights into the role of individual CPG variants in the pathogenesis of tumour types rarely associated with either of the relevant genes might be derived from LOH studies, assuming the relevant inherited cancer genes are tumour suppressor genes. Examples presented here however, show positive results in tumours that are characteristic of variants affecting the studied locus.

When performed on DNA from GISTs diagnosed in the mother-son diad with *SDHA* and *PALB2* variants, results were suggestive of a causative effect of the former but not the latter. A number of reports in the literature performed LOH analysis, often indicating a predominant role for one of the variants such as the other *BMPR1A*/*PMS2* case[316] and, perhaps surprisingly, in the breast cancer from an individual with *BRCA1* and *CHEK2* variants where loss of the wild type *CHEK2* allele was shown.[317] Predominance of one variant was suggested in some cases of *BRCA1/BRCA2* MINAS. For example, analysis of three primary breast cancers from one individual demonstrated LOH at *BRCA1* in one tumour and at *BRCA2* in the other two,[255] suggesting that there was no direct interaction between the two loci in the tumours. However, a seemingly conflicting result was obtained in another case report where LOH at both loci was demonstrated in an ovarian cancer from the same patient.[254]

Caution should be exercised in the interpretation of results from LOH analysis as they can be uninformative if the somatic mutational event ("second hit") is a single nucleotide variant, indel or promoter methylation of the wild-type allele (i.e. no LOH). Where LOH is seen, extensive chromosome aberrations occurring later in tumour development may theoretically lead to loss of the wild type allele without that event being significant in initiation.

Immunohistochemistry (IHC) studies may also be useful in tumours from MINAS cases as lack of staining for the protein product of the variant containing gene/s implies causality. IHC analysis in two breast cancers and one ovarian cancer from three individuals reported in the literature with *BRCA1/BRCA2* MINAS showed loss of both BRCA1 and BRCA2 immunostaining,[254,255,315] suggesting significance of both variants. Loss of staining for SDHB (indicating disruption of the succinate dehydrogenase complex) was shown in the *SDHA*/*PALB2* GIST cases and the colorectal carcinoma from the *BMPR1A*/*PMS2* case exhibited loss of PMS2 expression. One drawback of IHC is that it requires the development of a specific assay per protein or protein complex as opposed to LOH analysis that is applicable to any locus with the same sequencing technique. Furthermore, positive staining indicates the presence of a protein but not normal function. The use of mutational signatures to analyse tumours is in its infancy but represents a further potentially valuable method to delineate the relative contribution of multiple CPG variants if their mutagenic effects are distinct.

5.5.2 - Phenotypic manifestations combinations of genes containing variants

An interesting aspect of patients with MINAS is whether pathogenic variants in particular combinations of genes are associated with a more severe phenotype (e.g. earlier onset of cancer or cancer types that would be unexpected with one of the variants in isolation). A less severe phenotype is also feasible. The wide variety of combinations of individual pathogenic constitutional variants means that, with the exception of *BRCA1/BRCA2* combinations, information regarding observed phenotypic effects is limited.

Leegte et al[243] described 12 cases of combined *BRCA1/BRCA2* pathogenic variant cases and suggested that there was no evidence of increased severity whereas Heidemann et al[242] reported eight cases and suggested that a more severe phenotype was observed in two. Other reports have been on a smaller scale but cumulatively, 61 cases were identified in the literature, 56 of whom were female. 54 breast cancers were diagnosed in 43 of these individuals with a mean age at diagnosis for a first tumour at 40.3 years and for all breast cancer 41.3 years. 13 ovarian cancers were diagnosed in 10 individuals (all multiple tumours were synchronous bilateral) with a mean age at diagnosis of 49.2. The peak incidence age of breast cancer in *BRCA1* pathogenic variant carriers is 41-50 years with an equivalent figure of 51-60 years for *BRCA2*. Peak incidence of ovarian cancer for both genes is 61-70 years.[71] The ages at diagnosis noted in the *BRCA1*/*BRCA2* MINAS cases are therefore at the lower end of the peak for breast cancer and somewhat lower than that for ovarian cancer. This might suggest a synergistic effect of concurrent variants but the numbers of individuals remain small and the series as collated is subject to publication and ascertainment biases (e.g. over-representation of founder variants, which may be more penetrant). Only four cancers occurred in these cases that are not typical of variants in *BRCA1* or *BRCA2* but one of these was a colorectal cancer occurring at age 35 where microsatellite instability studies were normal and no loss of MLH1, MSH2 or MSH6 was demonstrated on IHC. This malignancy may therefore have been contributed to by the identified constitutional variants (bowel cancer had been diagnosed in the proband's father) but no further tumour studies were performed.

Other combinations of breast CPG variants have been described including *BRCA1*/*CHEK2* (n=4), *BRCA2*/*CHEK2* (n=3), *BRCA1*/*ATM* (n=2), *BRCA1*/*PALB2* (n=1), *ATM*/*CHEK2* (n=1) and *ATM*/*PALB2* (n=1). No atypical tumours or particularly early ages at diagnosis were noted in these individuals except for a patient with a combination of pathogenic variants in *BRCA1* and *PALB2* where multifocal breast cancer (Table 5.10), uterine leiomyomas and a meningioma were also diagnosed.[304]

TP53 variants can cause LFS, which leads to predisposition to various cancer types and is strongly associated with early onset breast cancer. They were noted in combination with variants in *CHEK2*, *PALB2* and *BRCA1* (1 occurrence each). The *TP53*/*CHEK2* case (see above) had a phenotype consistent with LFS. The *TP53*/*PALB2* individual had been diagnosed with early onset ovarian cancer, which is not typical for variants in either gene but LFS is associated with a wide variety of malignancies. The age of onset for the breast cancer (20 years) in the *TP53*/*BRCA1* case is low but cannot be interpreted as evidence of synergy between the two variants because LFS characteristically causes pre-menopausal breast cancer with breast screening recommended from a woman's early twenties. One report of constitutional deleterious *BRCA2* and *TP53* variants was identified in the literature where the individual concerned had been diagnosed with cutaneous malignant melanoma, breast cancer, ovarian cancer and colon cancer between the ages of 65 and 74 years.[292] In a mouse model where the orthologues of both of these genes are conditionally knocked out in epithelial tissues (to avoid embryonic lethality), a greater incidence and earlier onset of mammary and skin carcinomas was observed in comparison to mice where only *Trp53* or *Brca2* was conditionally knocked out, suggesting a synergistic effect in these tissues.[321] Though the mouse model is not directly comparable to the human status, four cancers had occurred in the case of *BRCA2/TP53* MINAS but all at relatively advanced age.

In addition to the case of *BRCA2*/*NF1* MINAS case reported here, a further combination of variants in a hereditary breast and ovarian cancer gene (*BRCA1)* and *NF1* was identified in a patient with cutaneous features of Neurofibromatosis type 1 and early onset (age 35) breast cancer.[306] The exhibited phenotype is consistent with independent expression of each variant but of note is the fact that *NF1* and *BRCA1* are both located on the long arm of chromosome 17. The presence of early onset breast cancer and Neurofibromatosis type 1 in the patient's mother along with both variants being found in the proband may suggest that the two altered genes were *in cis.* Such information has significant implications for genetic counselling of families where multiple pathogenic variants are identified though interestingly, the proband's brother who also had Neurofibromatosis type 1, did not carry the *BRCA1* variant suggesting a recombination event in the mother.

The second most frequently reported examples of specific MINAS were combinations of variants in genes predisposing to colorectal cancers.[299–303,316] Interestingly, severe phenotypes were noted in two patients with *APC/MLH1* pathogenic variant combinations with jejunal cancer seen in one case[299] and accelerated polyp progression in the other.[302] In the *BMPR1A*/*PMS2* case identified in the literature, a colorectal adenocarcinoma with loss of PMS2 staining on IHC was diagnosed in the apparent absence of colorectal polyps, suggesting a lack of *BMPR1A* variant penetrance. However, there was a strong family history of polyps (including in two children) and the level of investigation for polyps in the proband is not evident from the article.[316]

The phenotypic consequences of MINAS may be easier to interpret when the two genes involved are associated with dissimilar and narrow phenotypes. Most of the newly reported cases here fall into this category with phenotypes generally indicating an independent mechanism of action, that is, a phenotypic effect consistent with the presence of each variant in isolation. There was some suggestion of increased penetrance in the *SDHA*/*PALB2* cases and a possible atypical tumour (colorectal cancer) in the *FLCN*/*TP53* case but this cannot be confidently asserted.

In the literature, there are various reports of *BRCA1/BRCA2* pathogenic variants in combination with those in a mismatch repair gene (Table 5.10). In general, these have not demonstrated clear evidence of a synergistic effect on the severity or nature of the phenotype, although one reported case with deleterious *BRCA1* and *MLH1* variants had severe manifestations including endometrial, ovarian, clear cell renal and bilateral breast cancers diagnosed at age 39 years. Both breast tumours showed loss of the wild-type *BRCA1* allele but also showed absent staining of *MLH1* on IHC and loss of the wild-type *MLH1* allele. This suggests that both constitutional variants were significant to breast tumorigenesis in this patient. The high number of tumours and the development of early onset RCC (not usually associated with *BRCA1* or *MLH1* variants) suggests a possible synergistic effect.[288]

Reports of MINAS cases with other specific gene combinations only involve a single proband, although four individuals with *MEN1*/*RET* MINAS were reported in a single family with the authors concluding that more aggressive disease was not exhibited despite evidence for penetrance of both variants.[296] Pathogenic *PTEN* variants, which affect the PI3K/Akt signalling pathway[322,323] are reported in combination with those in *TP53*,[297] *APC*[319] and *SDHC*[298] with tumours characteristic of each variant observed in all three cases. A number of the tumours in the *PTEN/TP53* case were not typical of a variant in either gene and early onset of colonic polyps and paraganglioma were noted in the *PTEN*/*APC* and *PTEN*/*SDHC* individuals respectively. PTEN normally acts via Akt to down regulate MDM2 (and therefore increase p53 levels) in addition to its other roles[322,323] so this interaction may lead to a more severe phenotype. A case of MINAS involving pathogenic *FLCN* and *APC* variants has also been reported.[294] Typical colonic polyps and a colorectal cancer at age 28 occurred, as well as recurrent pneumothoraces and facial papules. The features are consistent with an independent mechanism, though the authors suggested that the *FLCN* variant might have enhanced the tumorigenic process given the observation that somatic mutational events affecting *FLCN* occur frequently in (microsatellite unstable) colorectal cancers.[272]

There are inherent ascertainment biases influencing which MINAS cases are present in the literature (and amongst the newly reported cases here) including more frequent analysis of combinations of particular genes, the range of phenotypes referred for testing and the restriction of analysed genes to

only those most strongly suggested by the tumour history or examination findings (e.g. cutaneous manifestations of cancer predisposition syndromes). Availability, or lack thereof, of sequencing of certain genes in some centres may also be a factor and is likely to have led to recognition of four *FLCN* MINAS probands in the centres contributing to the current analyses where this gene is tested frequently and is the subject of research studies. The appearance of certain CPG variant combinations may not simply be related to the population frequency of individual CPG variants and *in utero* death resulting from certain combinations might lead to a paucity of them being detected clinically. These biases are, in part, likely to be reduced by a more comprehensive genetic testing strategy made possible by cancer gene panels or whole exome/genome sequencing, which is likely to result in increased recognition of MINAS.

Increasing detection will inevitably lead to increased demand for accurate information on the likely phenotypic effect of particular variant combinations, in particular whether a more severe (i.e. a synergistic interaction) or even attenuated phenotype is to be anticipated rather than the variants having an independent effect. The MINAS cases described here are broadly indicative of an independent expression of both variants whereby the chance of necessary further tumourigenic events (e.g. second hits) is not greatly influenced by the other variant. In such a scenario, the probability of developing a cancer (due to either CPG variant) might be increased to a degree due to a greater variety of possible tumour initiating events but this might not be observable clinically.

Despite the general picture of independent effects, some individuals appear to show earlier age at diagnosis or unusual/more numerous tumours and in certain circumstances it may be prudent to expect that particular combinations of aberrant genes might result in a more severe phenotype. In practice, it is difficult to distinguish between these effects and incidental unrelated tumours but tumour studies can be helpful. There are a number of feasible mechanisms whereby a synergistic effect may ensue such as increased genomic instability leading a greater chance of necessary further tumourigenic mutation. Tumour development might be encouraged by compromised function of components of two tumour suppressive pathways (e.g. DNA repair and cell cycle regulation) or the loss of two components in a single pathway may lead to enhanced downstream aberrant signal. Two gain-of-function variants in proto-oncogenes might predict a more severe phenotype (though no reports of such cases were found) because, in contrast to tumour suppressor genes, the further event of somatic inactivation of a wild-type allele is not required to initiate tumorigenesis. An intriguing potential way in which MINAS might influence phenotype is the situation where an individual has pathogenic variants in two tumour suppressor genes that map to the same chromosome region. Loss of part of a chromosome harbouring both wild type alleles will result in a tumour that is homozygous null for both. This may have occurred in the *FLCN/TP53* case as these genes map to 17p11.2 and 17p13.1 respectively.

It is also feasible that in some situations, MINAS might lead to an attenuated tumour phenotype that is milder than if one of the pathogenic CPG variants was present in isolation. Clinically, these cases would be difficult to recognise because individuals would be less likely to present to clinical services. Where MINAS has been identified in an individual, it may not be possible to distinguish between attenuation conferred by MINAS and non-penetrance as cancer predisposition syndromes are usually not fully penetrant. If numerous further cases are uncovered by routine multigene testing strategies in future however, opportunities may arise to compare MINAS individuals with single CPG variant carriers to observe for differences in phenotypic severity. The most obvious mechanism by which MINAS might be protective against neoplasia is synthetic lethality. This phenomenon has been demonstrated through the efficacy of poly ADP ribose polymerase (PARP) inhibitors in cancers arising in carriers of pathogenic variants in *BRCA1* or *BRCA2*, which have often undergone a second hit affecting the wild type allele. Resulting dysfunction of double stranded break repair is compounded by inhibition of base excision repair by PARP inhibition and tumour cells are unable to tolerate the compromise of both processes.[106] In tumours from MINAS cases that have undergone a second hit at one of the variant containing genes therefore, it might be anticipated that haplo-insufficiency or a second hit at the other loci may render the clone untenable in some cases.

5.5.3 - Data sharing

There are myriad possible combinations of high penetrance CPG variants but conclusions as to their effect, as with many genetic conditions, are limited by small numbers. A useful resource with which to discern the effects of individually rare combinations and improve future management of patients with MINAS would be a reference database containing clinical, genetic and tumour information. Such information could guide the clinician as to what the effect of each combination of aberrant genes might be and prompt collation of individuals for further study. To facilitate sharing of such information, the author has established an online registry where cases can be uploaded via the Leiden Open Variant Database and identified by the phenotypic tag “MINAS” (http://databases.lovd.nl/shared/diseases/04296).

At present, clinical cancer genetics services remain predominantly focused on identifying a small range of CPG variants leading to risk that is amenable to mitigation strategies. Conceptualising MINAS, and indeed variants, in this manner may therefore be useful in the short to medium term but risks emphasising a false dichotomy between disease and non-disease-causing variants when a spectrum of risk may be a more accurate view. In an era of genomics and effective personalised medicine, the role of moderate to low penetrance variants and polygenic risk scores is likely to become more prominent. In the fullness of time, case sharing platforms might include a collection of risk conferring variants per individual in most cases although this may compound the issue of small

numbers of genetically similar individuals from which to draw conclusions regarding phenotypic effects.

Chapter 6 - Analysis for variants in putative novel loci associated with cancer predisposition genes in a multiple primary tumour series

Scripts used in these analyses are stored as an appendix in the form of a GitHub repository (https://github.com/jameswhitworth/Thesis-Elucidating_the_genetic_basis_of_multiple_primary_tumours-Scripts_appendix doi:10.5281/zenodo.1501206). They are denoted with the prefix "RA" (repository appendix) in the text in and in the repository.

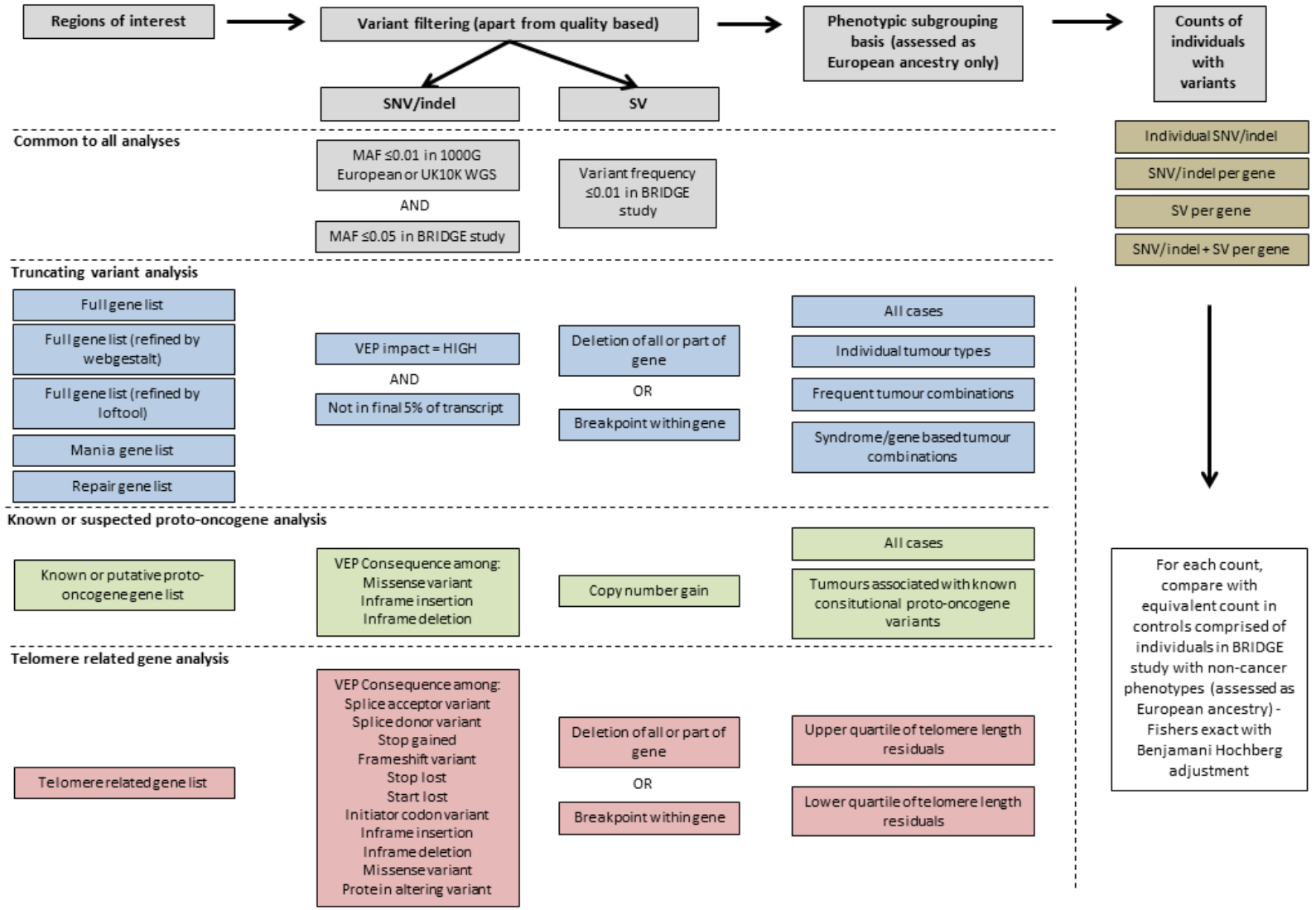
6.1 - Introduction

An aim of this project was to determine if studying individuals with multiple primary tumours (MPT) might lead to the identification of novel candidate loci relevant to cancer predisposition. To this end, data resulting from whole genome sequencing (WGS) of samples from MPT cases were used to perform case-control based analyses where the exposure of interest was the presence of a variant affecting loci of interest. A number of sets of genomic regions were proposed based on various lines of evidence suggesting a potential role in tumour susceptibility. The MPT series used for this purpose was that defined in Chapter 3 with some additional exclusions based on ancestry. A range of separate case control based studies were executed with the different loci of interest and phenotypic subdivisions of cases as described below.

Analyses were performed utilising counts of individuals with truncating variants affecting genes in lists that were compiled to include those that are recurrently mutated in somatic cancer studies, involved in DNA repair or functionally related to known cancer predisposition genes (CPGs) (section 6.2). Missense variants in known or putative proto-oncogenes causing tumour predisposition were also considered (section 6.3), as were coding variants affecting telomere related genes in individuals with estimated telomere length at the higher and lower end of that observed in the series (section 6.4). Frequency of variants in various non-coding regions was also analysed in cases vs controls (section 6.5). Regions of interest included enhancers and promoters of known CPGs (section 6.5.2.1) and ultra-conserved regions (section 6.5.2.2). Variants affecting expression quantitative trait loci (eQTL) reported to influence expression of CPGs in normal tissues (section 6.5.2.3) and cancer samples (section 6.5.2.3) were also counted and analysed. Many of the workflows used were common between the separate case-control based analyses. These are described in greater detail in the first section concerning the truncating variant analysis and subsequently referred to if used in other analyses. A summary of the study design is depicted in Figure 6.1 for coding variants and Figure 6.2 for non-coding variants.

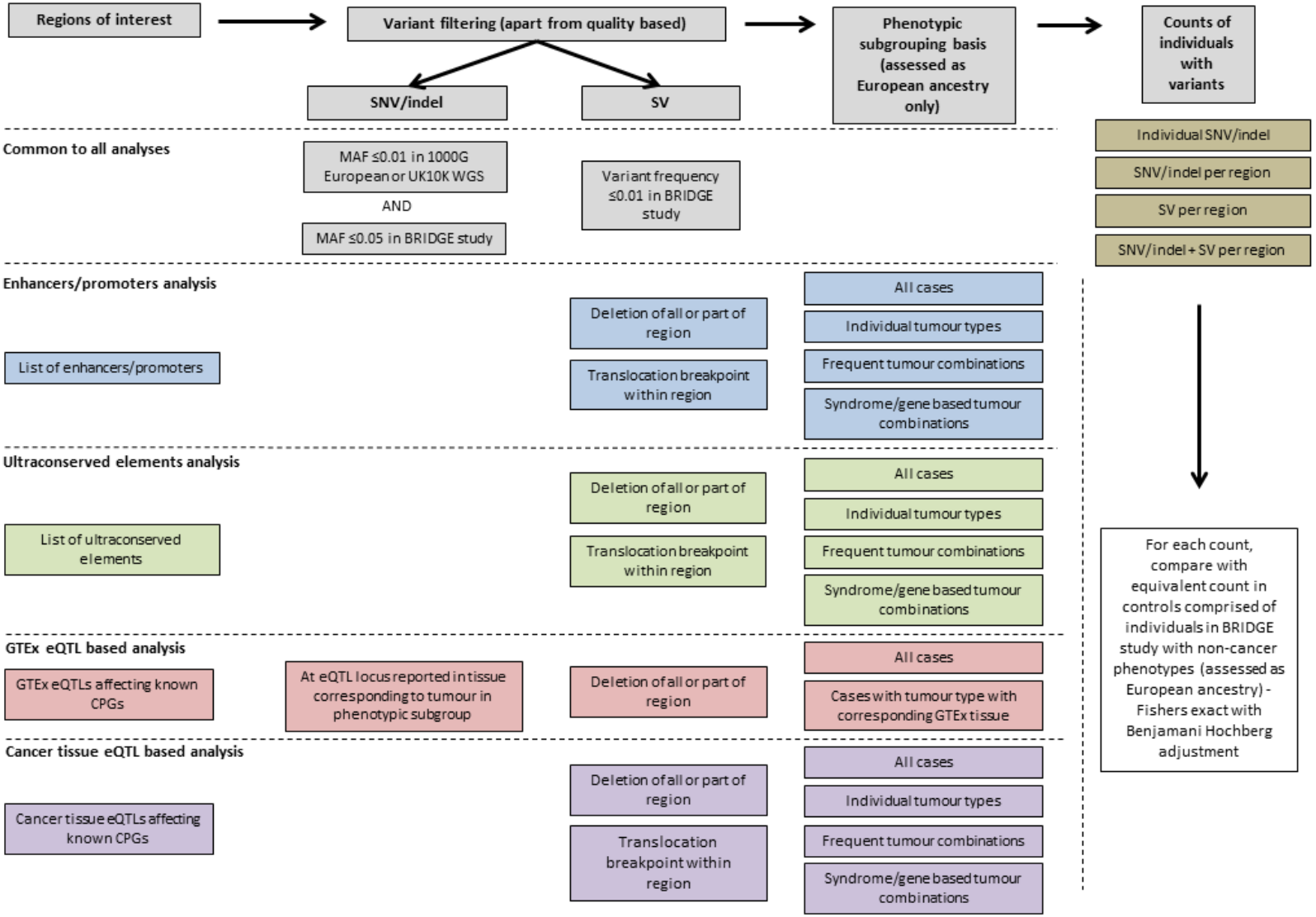
Additionally, a segregation-based analysis was performed on a family suspected to be manifesting a recessive cancer predisposition syndrome.

Figure 6.1 - Study design – Coding variants



MAF – Minor allele frequency, SNV – Single nucleotide variant, SV – Structural variant, VEP – Variant Effect Predictor

Figure 6.2 - Study design – Non-coding variants



CPG – Cancer predisposition gene, eQTL – Expression quantitative trait locus/loci, GTEx - Genotype Tissue Expression Project, MAF – Minor allele frequency, SNV – Single nucleotide variant, SV – Structural variant

6.2 Analysis of predicted truncating variants in known or suspected cancer predisposition genes

6.2.1 - Introduction

Sequencing studies to elucidate disease causing genetic factors often generate large numbers of potentially causative variants, particularly if massively parallel techniques such as WGS are used. The majority of these will not make a significant contribution to the disease phenotype in question and the classification of variants based on various lines of evidence is a large and active area of research. Assertions as to the pathogenicity of a variant are frequently made on the basis of predicted molecular consequence on the protein product and truncating consequences are generally regarded as strong evidence of a deleterious effect on function. Truncating variants include those inducing frameshifts, premature stop codons and aberrant splicing due to disruption of canonical splice sites. Such variants may lead to reduced protein function through absence of a functional portion of amino acid sequence or through nonsense mediated decay whereby premature stop codons within transcripts lead to detection and degradation by intracellular mechanisms following transcription.[324]

A degree of caution is necessary when assigning pathogenic status to a truncating variant as they have been shown to be frequent in individuals where no disease phenotype is evident. Previously, interrogation and variant assessment of data from 185 individuals in the 1000 Genomes project demonstrated around 100 truncating variants (including large deletions) per individual, around a fifth of which were homozygous, indicating complete inactivation.[325] A study of over 10,000 individuals from a society where consanguineous unions are common (Pakistan) and who were enrolled in a cardiovascular risk study demonstrated at least one gene with homozygous putative loss of function variants in 17.5% individuals.[326] Such occurrences are frequently termed “human knockouts.” Given findings such as these, it follows that mechanisms must exist whereby the presence of a constitutional truncating variant in an individual does not necessarily lead to disease. This may be the case even when the variant occurs in a gene known to be associated with medical conditions. *BRCA2* ENST00000544455 c.9976A>T (p.K3326*) is a nonsense variant that leads to truncation of the final 93 amino acids of the protein product. BRCA2 protein resulting from this transcript contains 3418 amino acids so this variant leads to a loss of less than 3% of the protein and BRCA2 function appears to be retained. The variant is generally regarded as benign and although there is some evidence to suggest an increased breast cancer risk associated with it, this is not at the level observed for other *BRCA2* truncations.[327] Truncating variants may not lead to the loss of large part of a protein product even if they occur at more 5’ locations within the gene. *TP53* ENST00000617185 c.387C>G (p.Y126*) occurs 215 amino acids before the end of the transcript but has been demonstrated to produce a full length protein with retained function through the generation of an alternative splice site.[328] A nonsense variant ENST00000357033 c.4250T>A (p.Leu1417X) in *DMD*, a gene associated

with Duchenne and Becker muscular dystrophy, has been observed in an individual with a phenotype intermediate between the two disease subdivisions. Reverse transcriptase polymerase chain reaction studies demonstrated that the variant led to alternative splicing and skipping of a single exon,[329] a phenomenon that has been synthetically emulated with therapeutic intent.[330] Many diseases with a constitutional genetic basis and which can be caused by truncating variants manifest in specific tissues, implying that there may be a compensatory mechanism in unaffected tissues. A study of expression of disease gene paralogues across multiple tissues recently showed that lower levels of paralogue expression are observed in tissues generally affected by variants in the gene corresponding to that paralogue.[57] If such a putative compensatory situation occurred across a broad range of tissues then it may account for non-penetrance of truncating variants. Additionally, extensive aberration of protein function due to a truncating variant may be protective against disease, as evidenced by truncating variants in *PCSK9* leading to reduced cardiovascular risk due to reduced binding to low density lipoprotein (LDL) receptors and consequent reduced circulating LDL levels.[331]

Despite these potential mechanisms of reduced or absent disease-causing effect of truncations, the majority of known pathogenic variants in CPGs are truncating in nature. Indeed, a search of ClinVar with the 83 gene names used for the WGS-based comprehensive CPG analysis described in Chapter 4 showed 8,784 variants classified as pathogenic or likely pathogenic with 2* evidence or higher, 7,659 (87%) of which were classified as frameshift, nonsense or splice site. Analyses to consider the frequency of these variant classes in MPT cases vs controls were therefore undertaken.

Studies associating genetic variants with disease are assisted by focusing on proposed loci at which causative variants may reside. Benefits include reduced use of analytical resources, a lower chance of false negatives resulting from application of correction for multiple hypothesis testing, and, in the presence of a possible significant result, the provision of further lines of evidence of causality other than the association itself. Proposition of candidate loci may be through various means including linkage analysis and identification of genes more likely to be relevant to the disease in question. According to the latter strategy, a number of gene lists were curated based on possible relevance to cancer predisposition and the frequency of truncating variants within these genes was recorded. Given that there is significant overlap between CPGs and genes observed to be recurrently mutated in somatic cancer sequencing studies,[45] a list was formulated based on top results from The Cancer Genome Atlas (TCGA) studies. Additionally, gene lists were produced based on the ratio of non-synonymous to synonymous variants in cancer tissues, evidence of a role in DNA repair and functional relatedness to known CPGs. Variant counts at these loci were then used to perform case-control analyses on various phenotypic sub-groups within the MPT series. Given that structural variants such as chromosomal deletions and translocations affecting genes of interest may also lead to absent or non-functional protein products, their frequency was also considered.

6.2.2 - Methods

6.2.2.1 - Gene lists

Five gene lists were compiled that contained known CPGs or genes hypothesised to be CPGs. The methods used for compilation are described below.

6.2.2.1.1 - Genes somatically mutated in cancer sequencing studies and known cancer predisposition genes

In order to identify genes in which variants are significantly over-represented in malignant tumours, study summaries from all available TCGA studies were downloaded from the cBioPortal data portal.[34] TCGA is a collaborative project to perform somatic sequencing on a wide variety of tumour types on a large scale. One study each from the Broad Institute, Michigan Centre for Translational Pathology, Memorial Sloane Kettering Centre were also downloaded. Study summaries each contain a list of genes that were noted to contain variants in the cancer type studied, along with the frequency at which variants were recorded. A number of factors may influence the frequency at which a given gene is mutated in a sample of sequenced tumour tissue including gene size, expression level (due to transcription coupled repair),[332] background mutation rate of the specific tumour type and the time point at which replication occurs during the cell cycle.[333] The MutSig tool is designed to highlight significantly mutated genes while taking account of these processes and has been applied to many of the TCGA studies appearing in cBioPortal. Output is expressed as a p-value where the null hypothesis is no difference in mutation frequency in a given gene between tumour and control tissue. Downloaded study summaries (n=41, Table 6.1) that included this measure were selected and any gene with MutSig $p < 0.01$ (n=902) used for the gene list.

Table 6.1 - Cancer sequencing datasets with MutSig assessment downloaded from cBioPortal

Name of study	Research group	Published dataset (if TCGA)	Samples
Adenoid cystic carcinoma	MSK		60
Adrenocortical carcinoma	TCGA		90
Bladder urothelial carcinoma	TCGA		130
Bladder urothelial carcinoma	TCGA	Yes	130
Brain lower grade glioma	TCGA		286
Breast invasive carcinoma	TCGA	Yes	993
Breast invasive carcinoma	TCGA		982
Breast invasive carcinoma	TCGA	Yes	507
Cervical squamous cell carcinoma and endocervical adenocarcinoma	TCGA		194
Cholangiocarcinoma	TCGA		35
Cutaneous melanoma	TCGA		345
Diffuse large B cell lymphoma	TCGA		48
Gastric adenocarcinoma	TCGA		289
Gastric adenocarcinoma	TCGA	Yes	289
Glioblastoma	TCGA	Yes	91
Glioblastoma multiformae	TCGA		290
Head and neck squamous cell carcinoma	TCGA		279
Head and neck squamous cell carcinoma	TCGA	Yes	279
Kidney chromophobe	TCGA	Yes	65
Kidney chromophobe	TCGA		66
Liver hepatocellular carcinoma	TCGA		198
Lung adenocarcinoma	TCGA		230
Lung adenocarcinoma	TCGA	Yes	230
Lung squamous cell carcinoma	TCGA		178
Pancreatic adenocarcinoma	TCGA		146
Phaeochromocytoma and paraganglioma	TCGA		183
Prostate adenocarcinoma	TCGA		332
Prostate adenocarcinoma	TCGA	Yes	333
Prostate adenocarcinoma	Broad		112
Prostate adenocarcinoma metastatic	MCTP		61
Renal cell carcinoma - clear cell	TCGA		417
Renal cell carcinoma - clear cell	TCGA	Yes	424
Renal cell carcinoma – papillary	TCGA		161
Sarcoma	TCGA		247
Testicular germ cell cancer	TCGA		155
Thyroid carcinoma	TCGA		405
Thyroid papillary carcinoma	TCGA	Yes	248
Uterine carcinosarcoma	TCGA		57
Uterine corpus endometrial carcinoma	TCGA	Yes	248
Uterine corpus endometrial carcinoma	TCGA		248
Uveal melanoma	TCGA		80

TCGA – The Cancer Genome Atlas, Broad – Broad Institute, MCTP – Michigan Centre for Translational Pathology, MSK – Memorial Sloane Kettering

Five of the top twenty UK incident cancers[177] did not have a corresponding TCGA study summary with the MutSig tool applied (acute myeloid leukaemia, colorectal, oesophageal, ovarian, myeloma). In these instances, the publication linked to the cBioPortal study of the relevant tumour type[334–340] (n=7) was retrieved and interrogated to find genes that the authors reported as significantly mutated (n=94). MutSig had been applied in six out of seven of the publications whilst one publication had used the Mutational Significance in Cancer suite of tools for a similar purpose. The same p-value threshold was used (where quoted) as for the cBioPortal study summaries.[341]

To incorporate known CPGs into the gene list and potentially demonstrate novel phenotypic manifestations, all genes appearing in a comprehensive review of CPGs[45] (n=114) or sequenced by the Illumina TruSight Cancer gene panel assay (Illumina Inc., San Diego, CA, USA) (n=94) were added. Additionally, published CPGs *NTHL1*[36] and *CDKN2B*[182] that didn't appear in either of the two lists were included.

Compilation of the sources above produced a list of 1,060 gene names, which were converted to Ensembl gene identifiers with Ensembl BioMart.[185] Where multiple identifiers existed for a single gene name, the identifier linked to the gene name on the Ensembl browser[184] was used (also used to select gene identifiers for all gene lists described below). This process resulted in a final gene list of 1,055 unique gene identifiers, referred to in results tables as the "Full" gene list (Table A3).

6.2.2.1.2 - Refinement of gene list

The utilisation of techniques to correct for multiple hypothesis testing used in these analyses (see below) may lead to an increased probability of false negative results with a higher number of tests performed. Consequently, an attempt was made to identify genes on the list that were most likely to be significant in tumour predisposition with the intention of producing a refined list where there was less chance of type two error. Two techniques were used to produce two separate refined lists for further analysis.

Predicted nonsense, frameshift or splice (loss of function) variants may be tolerated due to lack of haploinsufficiency for a given gene. Loss of function variants in genes that don't exhibit haploinsufficiency are consequently likely to be more frequent in populations. LoFtool[342] is a method that considers the per gene ratio of loss of function to synonymous variants in Exome Aggregation Consortium (ExAC) data[161] to produce a ranking of genes according to predicted tolerance to functional loss of one allele. Ensembl variant effect predictor[186] was used to annotate the original gene list and the quartile of scores predicting greatest intolerance were selected to produce a refined LoFtool-based list of 469 genes. This is referred to in results tables as the "Loftool" gene list (Table A4).

Despite tools such as MutSig to identify somatically mutated genes that contribute to the cancers in which they are found, many genes highlighted by cancer sequencing studies may not be functionally relevant to tumourigenesis. To identify genes on the original list which were most likely to be functionally relevant, the WebGestalt tool[343] was used to identify gene ontology (GO) terms[344,345] enriched among those assigned to those genes. WebGestalt was run twice for biological process GO terms and molecular function GO terms and the significantly over-represented terms from each enquiry (false discovery rate <0.05) noted. Any gene with at least one assigned GO term among these outputs was retained to produce a refined list of 617 genes (Table A5), which is referred to in results tables as the "Webgestalt" gene list.

6.2.2.1.3 - Other gene lists utilised

Genes that frequently contain somatic cancer driver variants may be identified through methods other than counting the number of variants per gene in a given cancer type. A recent study analysed 29 cancer types (7,664 samples) and observed the ratio of non-synonymous variants to synonymous variants per gene. On the basis that genes which tend to accumulate positively selected (at the somatic level) variants in tumours are likely to have an increased ratio, a set of 179 genes under positive selection (false discovery rate <0.05) was generated[7] (Table A6). This list was utilised directly for downstream analysis and is referred to as the "CGP" list in results tables after the Cancer Genome Project that produced the original publication.

Many CPGs are involved in DNA repair, including those that are most frequently tested clinically (*BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, *PMS2*). Consequently, a further gene list (Table A7) was utilised comprising all genes assigned with the DNA repair GO term (GO:0006281) (n=446) and is labelled as the "Repair" gene list in results tables.

Novel CPGs may also be uncovered through their interaction with existing ones. A final gene list was compiled by identifying interacting partners of known CPGs (n=133, comprised of all those appearing in a comprehensive review of CPGs[45] (n=114), sequenced by Illumina TruSight Cancer gene panel assay (Illumina Inc., San Diego, CA, USA) (n=94), *NTHL1*[36] and *CDKN2B*[182]). Interactions were found using the GeneMania platform[346] and a list of 142 genes produced (Table A8), which is referred to as the "Mania" gene list in results tables.

6.2.2.2 - Variant filtering – Single nucleotide variants (SNVs) and indels (Script RA6.1)

For all gene lists, the Ensembl canonical transcript identifier was selected for each Ensembl gene identifier by referencing gene-canonical transcript pairs provided by ExAC.[161] Canonical transcripts are defined according the following hierarchy: 1) Longest Consensus Coding Sequence (CCDS)[183]

translation with no stop codons, 2) Longest Ensembl/Havana merged translation with no stop codons, 3) Longest translation with no stop codons and 4) If no translation, longest non-protein-coding transcript.[184] Lists of transcripts were then used to obtain GRCh37 coordinates for the protein coding regions within them with Biomart.[185] Regions that were designated as within a patch region were excluded. Coordinates were then used to produce BED files +/- 5 base pairs for use in filtering of variant call format (VCF) files.

BED files were used in conjunction with bcftools view (version 1.4)[170] to extract variants in the corresponding regions that had FILTER PASS annotation from merged VCF files (one file per chromosome) containing variants called from NIHR BioResource Rare Diseases project (BRIDGE) WGS data (all sequenced individuals) and/or 1958 birth cohort exome sequencing data (see below). Variants from the latter dataset are subject to differences in frequency from BRIDGE data due to differences in sequencing coverage, variant calling or quality filtering and they were not used for downstream hypothesis testing. Filtered per chromosome files were then merged using bcftools concat and filtered with bcftools filter to exclude genotypes where read depth (DP) was less than 10, genotype quality (GQ) was less than 30 (corresponding to an estimated probability of the genotype call being incorrect of 1/1000) and variant allele fraction was less than 0.3.

The filtered merged VCF was then annotated with Variant Effect Predictor (VEP) version 90,[186] including the LOFTEE plugin[347] to specifically annotate predicted loss of function variants (nonsense, frameshift and splice site) with flags to indicate low confidence of functional loss of an allele as a result of the variant. Annotated variants were then filtered with the VEP filter script to include variants where impact was assigned as HIGH and where LOFTEE did not indicate that a variant occurred in the last 5% of the transcript. Further filtering was performed to remove variants with an allele frequency of >0.01 in the 1000 Genomes European[166] or UK10K[163] whole genome datasets. Variants were also excluded if they had an allele frequency of >0.05 across all samples in the BRIDGE project (n=9,424). The final merged VCF, containing genotype information for all BRIDGE samples pertaining to each filtered variant was read into R[178] (version 3.4.4) for further analysis.

6.2.2.3 - Identification of structural variant calls affecting genes of interest (Script RA6.2)

Selective filtering for truncating variants affecting genes of interest is based on the rationale that such variants are more likely to lead to an absent or non-functional protein product than other types of variant. However, structural variants (SVs) such as chromosomal deletions and translocations may also have this effect and are not identified by variant calling algorithms designed for SNVs and indels. WGS data gives the potential to identify SVs and the frequency of these aberrations predicted to affect the genes of interest were also recorded.

Files containing SV calls by Canvas or Manta (txt format) were initially filtered by the BRIDGE project to retain those that occurred at a frequency of less than 1% across all BRIDGE samples (n= 9,110) and were not associated with a flag introduced by Manta or Canvas indicating a low-quality call. Those files were subsequently filtered again with an R script to only retain variants fulfilling minimum quality criteria (GQ ≥30 for Manta, QUAL ≥30 for Canvas).

SV modalities that are most likely to disrupt protein function were considered and files containing calls for predicted deletions (separate files from Canvas and Manta), translocations (Manta), inversions (Manta) and insertions (Manta) were used. Interrogation of variants to elucidate those that affected genes of interest (each gene list described above was considered separately) was based on identifying variants where the predicted breakpoints contained or occurred within exon or gene start/end coordinates downloaded from Ensembl Biomart[184] (GRCh37 build). Translocations, inversions and insertions may exert deleterious effects due to breakpoints in non-coding regions of genes and for these SV modalities, coordinate files corresponding to the length of the gene were used. To avoid including purely intronic deletions amongst potentially pertinent SV calls, coordinate files corresponding to coding exons of genes of interest were used for analysis of deletions called by Canvas or Manta.

The conditions that a variant call was required to fulfil in order to be considered as affecting a region of interest are outlined in Table 6.2 and were executed using an R script. Manta annotation contained confidence intervals describing the range of bases surrounding the predicted breakpoint that are likely to contain the true breakpoint. These values can be utilised to produce genomic positions corresponding to the minimum start, maximum start, minimum end and maximum end of a given SV. These values were used in the identification of Manta called SVs affecting regions of interest.

Table 6.2 - Conditions used to identify structural variants

Structural variant modality	Coordinate file used	Conditions for structural variant call to fulfil
Deletion (Manta)	Exon	Max. start < exon start AND min. end > exon end OR Min. start > exon start AND max. end < exon end OR Max. start < exon start AND (min. end > exon start and max. end < exon end) OR Min. end > exon end AND (max. start < exon end AND min. start > exon start)
Deletion (Canvas)	Exon	Start < exon start AND end > exon end OR Start > exon start AND end < exon end OR Start < exon start AND (end > exon start AND end < exon end) OR End > exon end AND (start < exon end AND start > exon start)
Translocation (Manta)	Gene	Min. start > gene start and max. start < gene end OR Min. end > gene start and max. end < gene end
Inversion (Manta)	Gene	Min. start > gene start and max. start < gene end OR Min. end > gene start and max. end < gene end
Insertion (Manta)	Gene	Min. start > gene start and max. start < gene end OR Min. end > gene start and max. end < gene end

6.2.2.4 - Defining phenotypic groups

The multiple primary tumour cases participating in this study were phenotypically heterogeneous. Whilst some CPGs are associated with a diverse range of tumour types, variants in many of them are implicated in a narrower selection of neoplasms. For such CPGs, causative variants may be more readily detectable in a case-control study design where homogeneity of cases is enhanced. This is on the basis that in a situation where a particular set of variants cause a phenotype, signal is less likely to be diluted by an increase in N due to cases that don't conform to that phenotype. To this end, MPT probands were subdivided into 107 groups based on phenotype. Any proband being assigned to a subgroup had to first fulfil the general eligibility criteria of being diagnosed with two primary tumours under the age of 60 years or three under the age of 70 years. Additionally, only those assessed as European ethnicity were included to prevent misinterpretation of allele frequencies solely due to ancestral differences (individuals assessed as European ethnicity accounted for 424/452 MPT cases fulfilling general eligibility criteria).

The analysed subgroups, along with the number of individuals with each one, are outlined in Table 6.3 and included an analysis including all eligible participants. Single tumour subgroups contained all individuals diagnosed with a given tumour type before the age of 70 years. Combination subgroups (referred to as "2 from 2" in tables) were proposed by the presence of any discordant tumour combination (both before age 70 years) in an MPT individual. Subgroups which contained any individual diagnosed with ≥1 (or ≥2) of a selection of tumours (e.g. 1 from 2, 2 from 4 etc.) were also proposed and were intended to represent the tumour spectrum associated with existing cancer predisposition syndromes from literature review. This was based on the rationale that novel CPGs are often functionally related to existing ones. Two further subgroups were put forward due to identification of commonality of genomic aberrations (incorporating single nucleotide and copy number variants) across multiple tumour types in a large pan-cancer analysis based on TCGA data.[348] An R script was utilised to extract the sample identifiers for all individuals fulfilling the conditions required be included in a subgroup. To provide greater confidence in any forthcoming statistically significant results, phenotypic subgroups were only analysed if they contained at least three individuals in the case of single tumour groups or five individuals in any other group. A lower threshold was chosen in the single tumour groups given the suspected reduction in phenotypic heterogeneity vs other groups. For structural variant analysis, Canvas and Manta calls were only available for 360/424 individuals. Consequently, the sizes of phenotypic subgroups were reduced for any variant frequency comparison involving SVs and are outlined in Table 6.3.

Table 6.3 - Phenotypic subgroups used in analysis

No. individuals	No. individuals for counts incorporating structural variants	Syndrome, gene or phenotypic grouping forming basis of subgroup	No. tumours required to be included	Tumours
424	360	Nil	N/A - All MPT individuals	All
273	231	*STK11*	1 From 4	Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal
260	219	*PTEN*	1 From 3	Breast, Thyroid, Endometrium
253	216	Genomic commonality[41]	1 From 4	Breast, Aerodigestive tract, Lung, Ovary
241	206	*BRCA1, BRCA2*	1 From 2	Breast, Ovary
241	210	*TP53*	1 From 8	Breast, ACC, CNS, Soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma
219	188	*PALB2*	1 From 2	Breast, Pancreas
215	186	*CDH1*	1 From 2	Breast, Gastric
215	186	Nil	1 From 1	Breast
173	143	Lynch syndrome	1 From 4	Colorectal, Endometrium, Ovary, Sebaceous
141	117	Genomic commonality[41]	1 From 2	Colorectal, Endometrium
115		*BAP1*	1 From 6	Uveal melanoma, Kidney, Melanoma, Lung, Mesothelioma, CNS meningioma
99	82	*BMPR1A*	1 From 2	Colorectal, Gastric
99	80	*WRN*	1 From 7	Soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma, Thyroid
98	81	Nil	1 From 1	Colorectal
77	68	Xeroderma Pigmentosum	1 From 2	NMSC, Melanoma
77	68	*VHL*	1 From 4	Kidney, Phaeochromocytoma, Paraganglioma, CNS haemangioblastoma
74	64	*RMRP*	1 From 2	NMSC, Haematological lymphoid
74	64	*DOCK8*	1 From 2	NMSC, Haematological lymphoid
67	59	*RB1*	1 From 7	Retinoblastoma, Soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma
64	56	*TSC1*, *TSC2*	1 From 3	Kidney, Kidney angiomyolipoma, CNS
58	52	*FLCN*	1 From 4	Kidney, Adrenal oncocytoma, Kidney oncocytoma, Fibrofolliculoma
58	52	*FH*	1 From 4	Kidney, Uterine leiomyoma, Uterine sarcoma, Cutaneous leiomyoma
56	50	Nil	1 From 1	Kidney

53	46	*NBN*	1 From 3	Haematological lymphoid, CNS, Soft tissue sarcoma
52	45	*TERT*	1 From 4	Haematological myeloid, Aerodigestive tract, Anus, Melanoma
51	42	Nil	1 From 1	Endometrium
50	36	*CDKN1B*	1 From 2	Thyroid, Pituitary
50	41	*CDKN2A*	1 From 3	Melanoma, Pancreas, CNS
50	42	Nil	1 From 1	Ovary
47	39	*RECQL4*	1 From 2	NMSC, Bone sarcoma
45	38	Nil	1 From 1	NMSC
44	32	*PRKAR1A*	1 From 3	Cardiac myxoma, Thyroid, Ovary sex cord-gonadal stromal
44	36	*MEN1*	1 From 8	Pituitary, Parathyroid, ACC, GINET, Lung carcinoid, Ovary neuroendocrine, Paraganglioma, Phaeochromocytoma
44	39	*STK11*	2 From 4	Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal
43	35	*PTEN*	2 From 3	Breast, Thyroid, Endometrium
42	37	Nil	2 From 2	Breast, Colorectal
38	33	*GBA*	1 From 2	Haematological lymphoid, Haematological myeloid
38	34	Genomic commonality[41]	2 From 4	Breast, Aerodigestive tract, Lung, Ovary
38	27	Nil	1 From 1	Thyroid
37	30	Neuroendocrine tumours	1 From 6	GINET, Lung carcinoid, Ovary neuroendocrine, Paraganglioma, Phaeochromocytoma, PNET
36	32	Nil	1 From 1	Melanoma
33	30	Nil	1 From 1	Haematological lymphoid
32	29	*SDHA*	1 From 3	Phaeochromocytoma, Paraganglioma, GIST
31	27	Sarcomas	1 From 5	Soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma
28	24	Lynch syndrome	2 From 4	Colorectal, Endometrium, Ovary, Sebaceous
27	22	CNS tumour	1 From 4	CNS, CNS haemangioblastoma, CNS meningioma, CNS nerve sheath
27	21	Fanconi anaemia	1 From 5	Haematological myeloid, Aerodigestive tract, Oesophagus, Cervix, Penis
27	24	Nil	2 From 2	Breast, Endometrium
24	22	Nil	2 From 2	Breast, Ovary
23	21	*HRAS*	1 From 2	Soft tissue sarcoma, Bladder
23	19	*NF2*	1 From 3	CNS meningioma, CNS, CNS nerve sheath
21	19	Nil	2 From 2	Breast, NMSC
20	17	Nil	2 From 2	Breast, Haematological lymphoid

18	14	*BUB1B*	1 From 3	Wilms, Soft tissue sarcoma, Haematological myeloid
18	18	Nil	2 From 2	Breast, Melanoma
17	13	*DKC1*	1 From 3	Haematological myeloid, Aerodigestive tract, Anus
17	13	*TP53*	2 From 8	Breast, ACC, CNS, Soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma
17	16	Nil	2 From 2	Endometrium, Ovary
17	14	Nil	1 From 1	Lung
17	15	Nil	1 From 1	Prostate
15	11	Nil	2 From 2	Breast, Thyroid
15	14	Nil	1 From 1	GIST
14	13	Nil	2 From 2	Breast, Kidney
13	11	*NF1*	1 From 3	CNS, PNS nerve sheath, PNS nerve sheath benign
13	11	Nil	1 From 1	Soft tissue sarcoma
12	10	Nil	1 From 1	CNS meningioma
12	11	Nil	1 From 1	Colorectal polyps
12	9	Nil	1 From 1	Pituitary
11	9	*BAP1*	2 From 6	Uveal melanoma, Kidney, Melanoma, Lung, Mesothelioma, CNS meningioma
11	9	Nil	1 From 1	Aerodigestive tract
11	10	Nil	1 From 1	Paraganglioma
10	8	Nil	2 From 2	Breast, Lung
10	9	Nil	2 From 2	Colorectal, NMSC
10	10	Nil	1 From 1	Bladder
9	7	Nil	2 From 2	Breast, Soft tissue sarcoma
8	7	*RET*	1 From 2	Thyroid medullary, Phaeochromocytoma
8	6	Nil	2 From 2	Colorectal, Endometrium
8	3	Nil	2 From 2	Kidney, Thyroid
8	6	Nil	1 From 1	CNS
8	7	Nil	1 From 1	PNET
7	6	*CDC73*	1 From 2	Parathyroid, Bone benign
7	7	*WRN*	2 From 7	Soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma, Thyroid
7	6	Nil	1 From 1	Cervix

7	5	Nil	1 From 1	GINET
7	4	Nil	1 From 1	Pancreas
7	6	Nil	1 From 1	Phaeochromocytoma
6	5	Nil	2 From 2	Colorectal, Prostate
6	4	Nil	2 From 2	Colorectal, Thyroid
6	6	Nil	1 From 1	CNS nerve sheath
6	5	Nil	1 From 1	Testicular
5	5	Nil	2 From 2	Breast, Cervix
5	5	Nil	2 From 2	Breast, CNS meningioma
5	4	Nil	2 From 2	Kidney, Lung
5	3	Nil	1 From 1	Haematological myeloid
5	3	Nil	1 From 1	Lung carcinoid
5	4	Nil	1 From 1	Uveal melanoma
4	3	Nil	1 From 1	Bone benign
4	3	Nil	1 From 1	CNS haemangioblastoma
4	3	Nil	1 From 1	Ovary sex cord-gonadal stromal
4	4	Nil	1 From 1	PNS nerve sheath benign
4	4	Nil	1 From 1	Salivary gland
4	4	Nil	1 From 1	Small bowel
3	3	Nil	1 From 1	ACC
3	3	Nil	1 From 1	Kidney oncocytoma
3	2	Nil	1 From 1	Oesophagus
3	3	Nil	1 From 1	Parathyroid

ACC – Adrenocortical carcinoma, CNS – Central nervous system, GINET – Gastrointestinal neuroendocrine tumour, GIST – Gastrointestinal stromal tumour, NMSC – Non-melanoma skin cancer, PNET – Pancreatic neuroendocrine tumour, PNS – Peripheral nervous system

6.2.2.5 - Control group

In order to compare variant frequency in cases vs controls, a control group (n=4,053) was used that was made up of participants (assessed as European ethnicity) in other non-MPT arms of the BRIDGE project. Genotype data from these individuals was included in the merged VCF used for analysis of cases. This control dataset offered the advantage of having been sequenced and bioinformatically processed in an identical manner to cases, minimising the probability of observing differences in variant frequencies due to discrepancies in those processes between datasets. A disadvantage is that BRIDGE participants predominantly exhibit rare disease phenotypes, which may be caused by genetic variation that could also contribute to tumourigenic processes. To counter this, the recruitment criteria for different arms of the project were reviewed and samples excluded if they belonged to an arm where there was considered to be a higher probability of neoplastic processes occurring. Individual phenotypic information was not available to perform exclusions on a case by case basis. A summary of the constituent samples of the control set can be found in Table 6.4.

Table 6.4 - Control group derived from non-MPT arms of BRIDGE project

Acronym	Full name	No. samples	No. samples (European)	Rationale for exclusion
SPEED	Specialist pathology evaluating exomes in diagnostics	1389	869	N/A
PAH	Pulmonary arterial hypertension	1157	966	Variants in *BMPR2* can be causative. Bone morphogenetic protein signalling downregulated in some cancers.
PID	Primary immune disorders	1371	1078	N/A
BPD	Bleeding, thrombotic and platelet diseases	1170	984	N/A
GEL	Genomics England pilot	2000	1694	May include suspected inherited cancers. Unknown if case or control.
PMG	Primary membranous glomerulonephritis	193	167	N/A
SRNS	Steroid resistant nephrotic syndrome	252	166	N/A
ICP	Intrahepatic cholestasis of pregnancy	270	190	N/A
HCM	Hypertrophic cardiomyopathy	253	227	N/A
SMD	Stem cell and myeloid disorders	257	130	Fanconi anaemia phenotypes
CSVD	Cerebral small vessel disease	250	233	N/A
NPD	Neuropathic pain disorders	195	139	N/A

For structural variant analysis, Canvas and Manta calls were only available for 3,889/4,053 individuals and this was the size of the control group used for any comparisons involving structural variant frequency.

VCF files from the 1958 birth cohort exome sequencing data[349] were also interrogated for variants in the regions of interest though allele frequencies would only be used as a further line of evidence in the event of potentially significant results highlighted by other comparisons (rather than hypothesis testing). This is due to the high number of apparently spurious results generated due to differences in sequencing coverage and pre-VCF variant filtering.

6.2.2.6 - Variant counting and hypothesis testing – Single nucleotide variants and indels (Script RA6.1)

Counting of variants and hypothesis testing was performed in R. For each of the gene lists, the variant table read into the R environment was subset to only include variants within a gene on that list. Subsequently, for every phenotypic subgroup, frequency of each variant in cases and controls was recorded with separate counts generated for heterozygous and homozygous genotypes. To test for statistically significant differences in frequency of variants detected in cases vs controls, a contingency table was constructed for each variant to denote cases and controls with or without the variant. Separate tables were produced for heterozygous, homozygous and summed heterozygous and homozygous genotypes. A Fishers exact test[350] was then performed on contingency tables to test the null hypothesis of no difference in variant frequency in cases vs controls. This test was considered appropriate given that analysis was based on rare truncating variants and, for each test, it was expected that one of the values in the contingency table would be less than five. To allow for multiple hypothesis testing, Benjamani-Hochberg correction[351] was applied to the p-values generated from all hypothesis tests where the number of tests was taken as the number of variant sites present in the analysed variant table (including all BRIDGE individuals and 1958 birth cohort individuals). Rather than simply increase p-values as a direct function of the number of tests (as in more conservative methods such as Bonferroni correction), this technique takes into account the distribution of p-values generated by all the tests in the experiment to produce a false discovery rate expressed as a q-value. A q-value of 0.05 implies that amongst all p-values in the experiment with a q-value <0.05, 5% will be false positives.

Pathogenic truncating variants in a given CPG are often diverse and analysis of individual variants may not detect genes in which variants are over-represented in cases vs controls. Consequently, counts of individuals harbouring ≥1 variant in a given gene were also analysed. For each phenotypic subgroup, contingency tables were produced for every gene on the analysed gene list comprising counts of cases and controls with or without a variant in that gene. Individuals with heterozygous, homozygous and heterozygous or homozygous variants were considered separately. Fishers exact tests were again applied to the contingency tables with Benjamani-Hochberg correction where the

number of tests was taken as the number of genes on the analysed list. The process was repeated for each of the gene lists.

6.2.2.7 - Variant counting and hypothesis testing - Structural variants (Script RA6.2)

The "per gene" analysis was also undertaken using counts of structural variants. For each gene on the gene lists (as used for analysis of single nucleotide variants and indels), the number of individuals with a variant fulfilling one of the qualifying criteria (Table 6.2) was noted to produce counts for case and control groups. As individual structural variants are unlikely to be shared between unrelated individuals and given the margin of error in precise predicted breakpoints, counts of individual variants were not considered. Rather, comparisons were made of the frequency of individuals with a structural variant affecting a given gene in cases vs controls. The same phenotypic subgroups were used as for single nucleotide variant and indel analysis although the number of included individuals within these subgroups was frequently reduced due to variant calls pertaining to a number of individuals being unavailable.

6.2.2.8 - Variant counting and hypothesis testing – Single nucleotide variants and indels combined with structural variants (Script RA6.3)

Under the rationale that structural variants and single nucleotide variants/indels affecting a given gene may both lead to loss of a functional protein product, counts of variants were combined amongst cases and controls. For each gene on an analysed list therefore, the total number of individuals with a qualifying variant of any type could be compared with that observed in controls. As per structural variant analysis in isolation, the numbers of cases and controls included in these analyses were smaller due to structural variant calls for some individuals being unavailable.

6.3 Analysis of variants in known or putative proto-oncogenes

6.3.1 - Introduction

Whilst most described CPGs are tumour suppressor genes, a number of cancer predisposition syndromes are due to deleterious variants in proto-oncogenes. Such variants lead to tumourigenesis through a gain of function mechanism and are typically non-truncating. Examples include Multiple Endocrine Neoplasia Type 2, which is associated with susceptibility to medullary thyroid cancer, phaeochromocytoma and parathyroid tumours. Causative variants are missense and affect a relatively narrow range of codons of the *RET* gene, leading to dimerisation of the receptor tyrosine kinase gene product and/or persistent signalling.[13] *MET* encodes a further receptor tyrosine kinase where missense variants (in the tyrosine kinase domain) can cause Hereditary Papillary Renal Cell Carcinoma.[14] Studies to identify novel CPGs frequently (as above) prioritise frameshift, nonsense and splice site variants but this strategy will not detect potentially causative gain of function variants in proto-

oncogenes. Consequently, data resulting from MPT cases were also interrogated for missense variants and inframe insertions/deletions in genes with functional similarity to existing proto-oncogene CPGs. Counts of variants were used for further comparisons of cases and controls. Structural variant calls were also interrogated to identify SVs predicted to lead to increased copy number of these genes with subsequent use in a separate analysis as well as one combined with single nucleotide variants and indels.

6.3.2 – Methods

6.3.2.1 – Gene list composition

In order to compile a list of known and putative gain of function CPGs proposed by functional relatedness, a comprehensive review of CPGs[45] was interrogated to elicit any CPG annotated with a gain of function mechanism of action (*ALK*, *CDK4*, *EGFR*, *HRAS*, *KIT*, *MET*, *PDGFRA*, *PTPN11*, *RET*, *RHBDF2*, *SOS1*). Resulting genes were annotated with HUGO Gene Nomenclature Committee (HGNC) gene family identifiers and all terms pertaining to known gain of function CPGs were compiled and reviewed. All of these terms considered consistent with tumourigenic processes were used to search the HGNC Gene Families Index[352] (Table 6.5) and all gene names assigned with these terms downloaded for use in analysis (Table A9). Gene names included in the downloaded table (n=184) were used to obtain a list of canonical transcripts and coding region genomic coordinates as described for truncating variants.

Table 6.5 - HUGO Gene Nomenclature Committee gene families used to search for possible proto-oncogene CPGs

Identifier	Description
321	Receptor Tyrosine Kinases
496	Cyclin dependent kinases
1096	Erb-b2 receptor tyrosine kinases
389	RAS type GTPase family
812	Protein tyrosine phosphatases, non-receptor type
722	Rho guanine nucleotide exchange factors

6.3.2.2 – Variant filtering and case control comparison (Scripts RA6.4 , RA.6.5 and RA6.6)

Coordinates were used to extract variants from WGS VCFs as per truncating variant analysis. Following annotation of the resulting merged VCF with VEP, variants were filtered based on allele frequency as previously described and on consequence, with only variants annotated with the consequences “missense_variant”, “inframe_deletion” or “inframe_insertion” being retained. To identify predicted SVs causing increased dosage of the genes on the gene list, copy number gain variant calls called by the Canvas algorithm were interrogated. Only this modality of SV call was

considered as other variant types such as deletions and translocations (notwithstanding the possibility of fusion genes and displacement to more transcriptionally active sites) are less likely to be consistent with a gain of function mechanism. Files containing calls were searched as per truncating variant analysis but would only pass filters if the start of the variant call was at a genomic coordinate before the start of the gene and the end of the variant call was after the end of the gene as defined by coordinates downloaded from Ensembl Biomart.[184]

Variant counting and hypothesis testing were executed in the same manner as for the truncating variant analysis. Phenotypic subgroups used were restricted to the group containing all MPT cases (n=424) and another comprised of all individuals diagnosed with at least one tumour amongst a list of neoplasms known to be associated[45] with existing gain of function CPGs (Melanoma, Lung, Bladder, Gastrointestinal stromal tumour, Kidney, Thyroid medullary, Phaeochromocytoma, Paraganglioma, Soft tissue sarcoma, Haematological myeloid, n=152). The number of tests for Benjamani-Hochberg adjustment in analysis of individual variant frequency was taken as the number of unique variants detected in cases or controls. For counts of individuals with variants per gene, the number of tests was taken as the number of genes on the gene list (n=184).

For each gene on the gene list, the number of individuals with a copy number gain SV (Table 6.2) was recorded in case and control groups and hypothesis testing performed with those counts. SV counts were also combined with SNV/indel counts for each individual each gene and compared. The same two subgroups were analysed but the number of individuals in each was reduced due to non-availability of SV calls for some participants (All MPT cases n=360, 1 from 10 of Melanoma, Lung, Bladder, Gastrointestinal stromal tumour, Kidney, Thyroid medullary, Phaeochromocytoma, Paraganglioma, Soft tissue sarcoma, Haematological myeloid n=133). The control group was reduced to 3,889 individuals. The number of tests for multiple hypothesis correction was again the number of genes on the gene list (n=184).

6.4 Analysis of estimated telomere length and counts of variants in genes related to telomere function in individuals with multiple primary tumours

6.4.1 -Introduction

Telomeres are repetitive sequences located at the ends of chromosomes that have a role in avoidance of genomic instability that may ensue through recognition of chromosome ends as areas of DNA damage. The process of cell division leads to a shortening of telomeres due to incomplete synthesis of the lagging strand by polymerases and further processing of chromosome ends to maintain telomere structure.[353] It follows that ageing should be associated with shortening and this has been observed in a number of studies. A systematic review of length measurement studies estimated the rate to be

around 20 base pairs per year.[354] Regulation and maintenance of telomeres is executed by two primary complexes. Telomerase lengthens them by adding repeats through a reverse transcriptase mechanism but has reduced activity in human tissues after embryonic development. Shelterin binds to telomeres and has a role in regulating telomerase activity as well as inhibiting DNA damage responses such as ATM activation and non-homologous end joining.[353]

Telomere maintenance is known to be relevant to the development of cancer but observations relating to telomere length in tumour and germline samples from individuals with neoplasia have led to a complex picture. As telomeres become shorter, they may become more vulnerable to DNA repair mechanisms that lead to chromosome aberrations. Resulting genome instability can potentially lead to somatic changes necessary for tumour development and both shortened telomeres and chromosome abnormalities indicative of unprotected telomeres have been observed in studies of cancer.[355–357] Furthermore, constitutional pathogenic variants in the telomerase reverse transcriptase component gene *TERT* are associated with predisposition to particular cancers and affected individuals have been demonstrated to exhibit shorter telomere length.[358,359] Familial pulmonary fibrosis is associated with an increased risk of lung cancer whereas Dyskeratosis Congenita, also associated with *TERT* variants, causes nail dysplasia, oral leukoplakia and cutaneous pigmentation abnormalities as well as predisposition to acute myeloid leukaemia and aerodigestive tract cancers. Shorter telomeres have also been observed in *BRCA1* and *BRCA2* pathogenic variant carriers vs controls.[360]

Despite observations such as these indicating an association with shorter telomere length and neoplasia, most human cancers show upregulated telomerase activity.[361] A constitutional variant in the *TERT* promoter that causes upregulation of telomerase has been identified in a family with multiple occurrences of melanoma and subsequently observed recurrently in melanoma cell lines from sporadic cases.[362] Constitutional loss of function variants in *POT1*, part of the shelterin complex, have also been seen in familial melanoma cases and shown to reduce binding to telomeres.[363] Affected individuals had longer telomere length, seemingly related to the normal role of *POT1* in inhibiting telomerase activity.[353] Additionally, a large study of around 95,000 individuals demonstrated an association between genetic determinants of longer telomere length (three single nucleotide polymorphisms in telomerase component genes) and increased cancer risk. Increased risk of lung cancer and melanoma was also shown to be associated with longer telomere length.[364]

To regard association of cancer with both shorter and longer telomere length/increased telomerase activity as contradictory would be to over-simplify interpretation of these phenomena. Telomere shortening can be regarded as a tumour suppressive mechanism as it is associated with activation of DNA damage responses, reduced proliferation and apoptosis. However, it may also increase the chance of chromosomal instability and tumourigenic aberrations. Subsequently, acquisition of

telomerase activity in cells that had sufficiently short telomeres to provoke these events could lead to developing tumour cells continued viability. Hypotheses to explain why longer telomeres can prompt cancer include avoidance of the tumour suppressive effects of telomere shortening and a dysregulated telomere phenotype with longer, unprotected telomeres.[353] Under the former model, tumourigenic mutational processes would predominantly be through means unrelated to shortened telomeres. In the latter model, oncogenic abnormalities would include telomere related structural variants such as telomere containing chromosome fusions, as have been observed in chronic lymphocytic leukaemias associated with somatic *POT1* variants.[365]

WGS provides an opportunity to gain insight into telomere biology through the estimation of their length in a DNA sample. This is due to the fact that, in contrast to targeted sequencing approaches, reads from telomeric regions are generated in sufficient numbers. Given that telomere length has relevance to tumourigenic processes, telomere length was estimated in MPT cases as well as controls. A regression model was fitted to estimated length vs age at sampling to assess deviation from the model. Both longer and shorter telomeres have been noted in individuals with cancer predisposition syndromes and within the MPT cases, two groups were identified who had length estimates within the top and bottom quartiles of residuals. Two case-control based analyses were then performed using the two groups as cases and regarding counts of variants in telomere related genes as the exposure of interest.

6.4.2 -Methods

6.4.2.1 - Analysing telomere length in BRIDGE BAM files (Script RA6.7)

To estimate leukocyte telomere length in individuals using WGS data, the Telomerecat package (version 3.2)[366] was used. This tool isolates sequencing read pairs from BAM files that are consistent with telomeric origin (contain ≥2 CCCTAA or TTAGGG sequences) to produce a "telbam" file. Telomere length is then estimated from the ratio of entirely telomeric read pairs to read pairs arising from telomeric and non-telomeric regions (as longer telomeres are more likely to produce read pairs entirely sequenced from telomeric areas). Telbams were produced for WGS BAM files with the telomerecat bam2telbam function and length estimates from telbams were generated with the telbm2length command. Prior to categorising telbam reads, telbam2length considers sequencing errors which involves generating a distribution of genotype quality scores from random loci and comparing it with a distribution from loci that are apparent mismatches to telomeric sequence. Consequently, non-identical outputs are generated from each run of telbam2length. To allow for this, ten outputs were generated for each telbam and the mean taken as the telomere length estimate for that sample.

6.4.2.2 - Estimated age at sampling

Telomere length reduces with cell division and is inversely correlated with age.[354] Measurement of it in this context should take into account the age at sampling. Documentation of this was provided by BRIDGE. For samples in the MPT arm (labelled MPMT in the BRIDGE project), the medical record was further reviewed to provide a date (or year if date not available) that the sample was taken. Samples were excluded from further analysis if an age was unavailable. A table was then compiled that linked per sample estimated telomere length with age at sampling. There is some evidence that telomere length is associated with ancestry[367,368] and non-European samples were excluded given that this would produce a minor reduction in the number of MPT cases and also that only European ethnicity samples would be used for variant frequency analysis downstream. It has previously been suggested that sex also influences telomere length but a meta-analysis to investigate this was not conclusive[369] and samples from male and female study participants were considered together.

6.4.2.3 - Fitting a linear model to estimated telomere length vs age at sampling and calculating residuals (Script RA6.7)

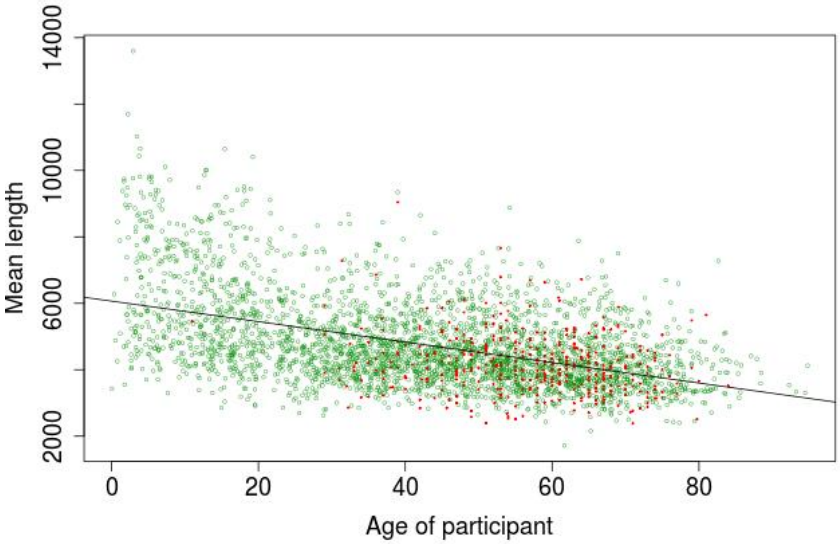
In order to assess the degree of deviance from expected telomere length given the age at sampling, the R lm function was used to fit a linear model (Figure 6.3) to the relationship between mean estimated telomere length and age at sampling across all 3,557 samples (Table 6.6, MPT, n=417 and non-MPT, n=3140). Significance testing of the model (F-statistic p-value < 2.2e-16) indicated rejecting the null hypothesis of no relationship between the variables.

Next, residuals based on the linear model were taken to provide a measure of how far the mean estimated telomere length deviated from the expected value for each individual given the age at sampling (Figure 6.4). Residuals of MPT cases were compared with non-MPT controls (Figure 6.5) with a Welch t-test (as Bartlett test indicated unequal variance between the groups, p = 3.371e-12), which showed significantly lower (i.e. shorter telomere length) residuals in the MPT group (p = 0.001105).

Table 6.6 - BRIDGE samples used in telomere length analysis

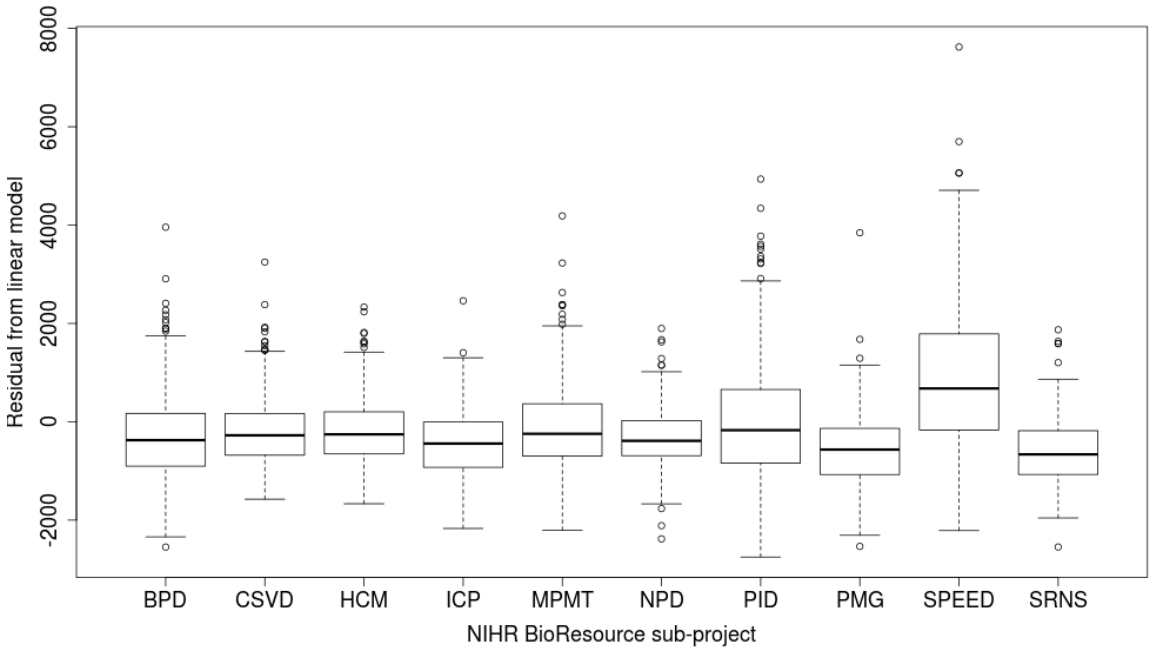
BRIDGE sub-project	Mean age at sampling	Number of samples
BPD	42.59351	481
PID	43.21670	1003
SPEED	34.54710	672
PMG	39.33607	150
MPMT	56.41230	417
ICP	35.94946	147
HCM	59.22851	188
CSVD	59.28183	197
NPD	51.53243	136
SRNS	33.58158	166

Figure 6.3 - Plot of linear model. MPMT individuals indicated by red points



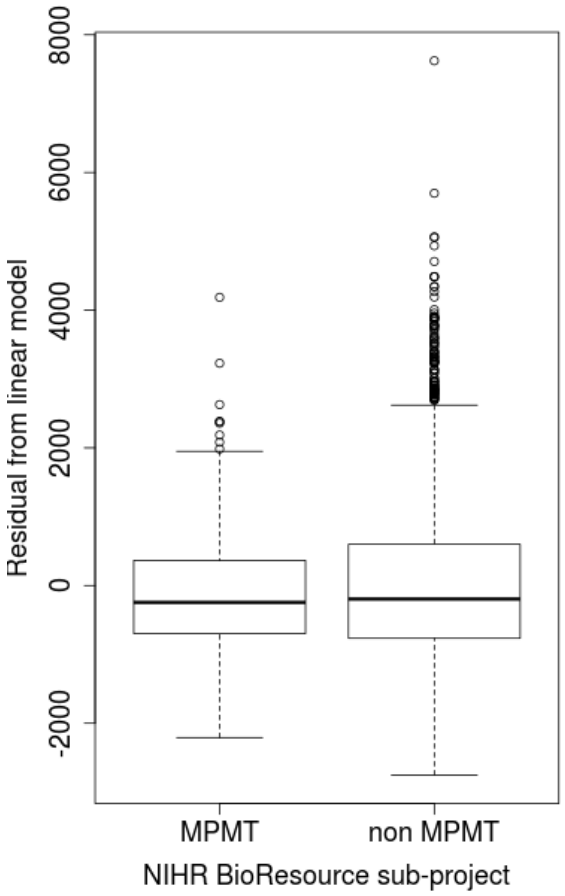
Red points indicate MPT samples. Green points indicate non-MPMT samples

Figure 6.4 - Plot of residuals by project



NIHR (National Institute of Health Research) BioResource also referred to as BRIDGE

Figure 6.5 - Plot of residuals MPMT vs non-MPMT



6.4.2.4 – Results of comparison of residuals between BRIDGE projects with discussion

Residuals as a function of telomere length were significantly lower (i.e. suggesting shorter telomeres) in the MPMT arm cases although the difference was not large. Extent of deviation from the linear model is susceptible to inaccuracies surrounding the documented date of sampling and this was not uniformly clear due to the fact that a large number of DNA samples were not from blood freshly taken for the purposes of the study. Furthermore, dates for non-MPMT BRIDGE arms could not be reviewed as part of the present analysis and may have been subject to biases. A large contributor to the difference in residuals between non-MPMT and MPMT appears to be the Specialist Pathology: Evaluating Exomes in Diagnostics (SPEED) study, which recruits paediatric cases with suspected monogenic neurological disorders. Any over-estimate in the age at sampling in that study could have led to the higher residuals. Alternatively, a poorer fit of the linear model at lower age at sampling is suggested by the scatter plot and could have contributed to greater deviations. A further possible explanation for comparatively shorter telomeres in the MPMT arm includes the effect of chemotherapeutic agents, which many participants would have been exposed to prior to blood sampling for DNA extraction. A study of 260 sporadic breast cancer patients treated with first line chemotherapy showed shorter telomere length than in controls, an effect that was also observed in 236 familial breast cancer cases. In both series, recovery of telomere length was also observed.[370] In a review of studies regarding the effect of a wide variety of chemotherapy drugs on telomere length in cell lines, a large majority of reports observed shortening.[371]

6.4.2.5 – Analysis of variants in telomere related genes amongst multiple primary tumour cases with shortest and longest residuals

To investigate the hypothesis that MPT cases with shorter or longer telomeres may have been predisposed to developing tumours due to a constitutional genetic variant in a telomere related gene (according to a list defined below), two case control analyses were performed where cases were identified by telomere length. To this end, the bottom and top quartile of residuals from the linear model in MPT cases were taken and corresponding individuals used to form a case group (n=107 for low residual group, n=105 for high residual group). The control group was made up of the same 4,053 European individuals used in the truncating variant analysis.

6.4.2.6 – Collating a list of telomere related genes

The variants of interest for analysis were those within genes documented as being related to telomere function. To formulate a gene list, the Gene Ontology database annotation file (version 2.1)[344] was downloaded and any line containing the character string “telomer” extracted. All GO terms within these lines were reviewed and a list of relevant terms compiled. Additionally, terms on this list were entered into the European Bioinformatics Institute QuickGo tool for GO term searches[372] to generate

an ontology term map. Any additional telomere related terms connected to the existing ones were then added to the list of terms of interest (Table 6.7). This amalgamated list of 19 GO terms was used to search the Gene Ontology annotation file to extract all gene names annotated with at least one of the terms (n=137) (Table A10).

Table 6.7 - Gene ontology terms relating to telomere function

Identifier	Description
GO:0003720	telomerase activity
GO:0010833	telomere maintenance via telomere lengthening
GO:0032204	regulation of telomere maintenance
GO:0032205	negative regulation of telomere maintenance
GO:0032206	positive regulation of telomere maintenance
GO:0032210	regulation of telomere maintenance via telomerase
GO:0051972	regulation of telomerase activity
GO:1904356	regulation of telomere maintenance via telomere lengthening
GO:1904357	negative regulation of telomere maintenance via telomere lengthening
GO:1904358	positive regulation of telomere maintenance via telomere lengthening
GO:0032211	negative regulation of telomere maintenance via telomerase
GO:0051973	positive regulation of telomerase activity
GO:0032212	positive regulation of telomere maintenance via telomerase
GO:0005697	telomerase holoenzyme complex
GO:0070034	telomerase RNA binding
GO:0000723	telomere maintenance
GO:0032201	telomere maintenance via semi-conservative replication
GO:0007004	telomere maintenance via telomerase
GO:0042162	telomeric DNA binding

6.4.2.7 – Variant filtering and case control comparison (Scripts RA6.8, RA6.9 and RA6.10)

As previously described for truncating variant analysis, gene names were used to identify canonical transcripts, Ensembl gene IDs and coding region genomic coordinates. A BED file based on these coordinates was used to extract variants in the regions of interest from WGS VCFs. The resulting merged VCF was annotated and filtered based on allele frequency as previously but all of the following consequence annotations could be included: “splice_acceptor_variant”, “splice_donor_variant”, “stop_gained”, “frameshift_variant”, “stop_lost”, “start_lost”, “initiator_codon_variant”, “inframe_insertion”, “inframe_deletion”, “missense_variant” or “protein_altering_variant.” Files containing SV calls were interrogated in the same manner as for truncating variants and any call fulfilling the filtering criteria was used to inform the counts of individuals with an SV predicted to affect each gene on the gene list.

Counting of variants and individuals with variants per gene with hypothesis testing was performed as per truncating variant analysis. The number of tests for Benjamani-Hochberg adjustment in analysis of individual variant frequency was taken as the number of unique variants detected in cases or

controls. For counts of individuals with variants per gene, the number of tests was taken as the number of genes on the telomere related gene list (n=137).

Counts of individuals with SVs per gene and SVs combined with SNVs/indels per gene were also considered as per truncating variant analysis. For these purposes the number of individuals in each group was reduced due to SV calls for some participants being unavailable (low residual group n=80, high residual group n=81). The size of the control group was reduced to 3,889. The number of tests for multiple hypothesis correction was again the number of genes on the gene list (n=137).

6.5 Analysis of non-coding variants potentially relevant to cancer predisposition

6.5.1 - Introduction

A key potential advantage of WGS in identifying constitutional variants predisposing to neoplasia is the ability to sequence non-coding regions. Although coding regions make up a small minority of the human genome, the majority of disease associated variants are within them.[373] Potential contributing factors to this observation are lower functional redundancy in coding regions and a hitherto restricted ability to sequence non-coding areas with assays commonly used in research studies.

The use of WGS in genetic research is increasing but the identification of individual non-coding variants that can cause Mendelian disorders has been infrequent. This is partly due to the difficulty in annotating non-coding variants with information that guides whether it is relevant for disease or not. Non-coding variants impacting on CPG function are consequently few in number but have been described. One example is *CDKN2A* ENST00000304494 c.-34G>T, which is within the 5' UTR, has been reported to disrupt splicing,[374] and has pathogenic status in ClinVar. Efforts to combine germline DNA sequencing and RNA sequencing in tissues have produced association of non-coding variants with gene expression[375] and may contribute to the elucidation of disease-causing variants.

Although the number of specific non-coding variants associated with cancer predisposition syndromes are low in number, a large body of evidence has accumulated that indicates which regions are more likely to be significant in disease causation. The ENCODE project is a notable accumulation of such evidence, which compiles the findings of a large number of experiments performed using a wide variety of assays.[216] An example is co-immunoprecipitation sequencing (ChIP-seq), which identifies regions of the genome bound to defined proteins of interest (e.g. proteins known to bind to DNA) through antibody binding to those proteins, pull down and subsequent massively parallel sequencing.[376] A further assay type utilised by the project identifies less condensed areas of chromatin (i.e. more likely to be transcribed) by their sensitivity to cleavage by DNAse enzymes.[377] Efforts such as ENCODE have resulted in a canon of non-coding regions where transcription or binding influence

gene expression including promoters and long range regulatory elements such as enhancers. Additionally, functional relevance of non-coding regions can be indicated by conservation across species and lists of these regions have been curated.[378]

Given that WGS data generated as part of the present study gave the opportunity to search for non-coding variants in regions potentially relevant to tumour development, frequency of variants affecting a range of such regions were recorded and compared with controls in a similar manner to the case-control based analyses described earlier in this chapter.

6.5.2 - Methods

Study design relating to non-coding variants is summarised in Figure 6.2.

6.5.2.1 - Enhancers and promoters (Scripts RA6.11, RA6.12 and RA6.13)

Non-coding regions of the genome may exert a phenotypic effect by affecting gene expression. Two recognised mechanisms involve promoters, which lie close to the genes whose transcription they influence, and enhancers, which are more distant.[379]

In order to identify promoters and enhancers which may affect CPGs, the GeneCards[380] database was searched with the gene names (n=133) corresponding to all genes appearing in a comprehensive review of CPGs[45] (n=114) or sequenced by the Illumina TruSight Cancer gene panel assay (Illumina Inc., San Diego, CA, USA) (n=94). Additionally, published CPGs *NTHL1*[36] and *CDKN2B*[182] that didn't appear in either of the two lists were included. For each page corresponding to an individual gene name (searched 07/09/2017), available information regarding relevant enhancers and promoters was extracted and reviewed to produce a list of regions of interest.

Reported promoters for a gene in GeneCards are based on the Ensembl database and expressed as Ensembl regulatory region identifiers.[184] All such identifiers (n=73) on the interrogated gene pages were taken and converted to GRCh37 coordinates with BioMart.[185] Enhancers associated with a gene in GeneCards (collated by the GeneHancer database[381]) are taken from a number of sources including the Encyclopaedia of DNA elements (ENCODE),[216] Ensembl, Functional Annotation of the Mammalian Genome (FANTOM5)[382] and VISTA,[383] a browser containing experimentally validated non-coding elements with enhancer activity. Putative enhancers are given "Elite GeneHancer" status if they are supported by ≥2 of these evidence sources and only these (n=1050) were taken for further use. Genomic coordinates for enhancers were obtained via download from the GeneCards website.

Coordinates corresponding to all elements of interest (n=1,123) were compiled and used to produce a BED file. This was in turn used to interrogate BRIDGE WGS data and produce a variant table with

filtering for quality and allele frequency as described for analysis of truncating variants. No filter was imposed for molecular consequence.

Files containing structural variant calls were also interrogated to identify variants predicted to disrupt any of the elements of interest using the same genomic coordinates as used in the BED file and the same quality and variant frequency filtering criteria used for truncating variant analysis. Consequences of SVs in non-protein coding regions are less readily predictable than for coding regions and only deletions (Canvas or Manta calls) or translocations (Manta calls) were considered further as they were considered to be more likely to cause functional disruption.

To assess for significant differences in frequency of variants within the non-coding regions of interest, variant counts and hypothesis testing (Fishers exact tests with Benjamani-Hochberg correction) was also performed as per the analysis of truncating variants. Frequency of each observed variant was considered where the number of tests (for correction purposes) was equal to the number of unique variants observed in cases or controls. Counts of individuals with variants in each of the non-coding elements were also analysed where the number of tests was the number of elements considered (n=1,123). The phenotypic subgroups used were the same as for truncating variant analysis. Counts of individuals with SVs and SVs combined with SNVs/indels in each element were also compared in cases vs controls. Reduction in the size of case and control groups due to SV call availability was as per truncating variant analysis.

6.5.2.2 - Ultra-conserved elements (Scripts RA6.14, RA6.15 and RA6.16)

Functional activity of non-coding regions can also be suggested by evolutionary conservation and further regions to analyse in MPT cases were identified in this way. The Database of Ultra-conserved Non-coding Elements (UCNE)[378] has curated 4,351 non-coding regions that exceed 200 base pairs in length and have ≥95% sequence homology between human and chicken based on data downloaded from the University of California Santa Cruz (UCSC) browser. Most are predicted to regulate transcription and are categorised as intergenic (n=2,139), intronic (n=1,713) or untranslated regions (n=499). Human hg19 UCNE data (downloaded 21/9/2018) was used to provide genomic coordinates for all reported elements. Using these coordinates, analysis of frequency of variants (SNVs/indels, SVs, SNVs/indels combined with SVs) in cases vs controls was performed in the same way as for enhancers and promoters.

6.5.2.3 - Expression quantitative trait loci (Scripts RA6.17, RA6.18 and RA6.19 for expression quantitative trait loci from Genotype Tissue Expression Project. Scripts RA6.20, RA6.21 and RA6.22 for expression quantitative trait loci from cancer tissue studies)

Association of non-coding variants with gene expression levels across multiple tissue types has recently been reported in two major publications.[375,384] Such variants have been termed expression quantitative trait loci (eQTL) and given that they may affect expression of CPGs, their role in the MPT series was also investigated.

The first set of eQTL considered were those identified by the Genotype Tissue Expression Project (GTEx) that were reported to affect expression of 83 CPGs appearing in the gene list used for the WGS-based comprehensive CPG analysis described in Chapter 4 and listed in Table 4.1. Genes on this smaller list of CPGs were considered to have more robust evidence for a role in predisposition to adult onset tumours. Variant-gene pairs reported by GTEx have been relatively recently described in a single analysis and the smaller list was utilised to provide greater confidence of phenotypic relevance in any potentially significant results observed. GTEx recently reported 12,546 unique variant gene-pairs (observation of the same pairs in multiple tissues meant that 48,452 variant-gene-tissue combinations were reported) from the analysis of 10,294 samples from post-mortem donors between the ages of 21 and 70 years.[375] Donors had never been diagnosed with metastatic cancer and had not been treated with chemotherapy or radiotherapy in the two years prior to death. All variant gene pairs containing observations from all 48 tissue types (Table 6.8) were downloaded from the GTEx portal (version 7). Those quoted as significant by GTEx (q value <0.05) and reported to affect the expression of a gene on the gene list were selected but excluded if the data indicated that an eQTL had a positive effect on tumour suppressor gene expression or negative effect on proto-oncogene expression.

Table 6.8 - GTEx tissue types

GTEx tissue	Tumour in participant prompting interrogation for variant-gene pairs observed in GTEx tissue
Adipose_Subcutaneous	Lipoma
Adipose_Visceral_Omentum	N/A (no tumours in series in this tissue)
Adrenal_Gland	Phaeochromocytoma, ACC
Artery_Aorta	N/A (no tumours in series in this tissue)
Artery_Coronary	N/A (no tumours in series in this tissue)
Artery_Tibial	N/A (no tumours in series in this tissue)
Brain_Amygdala	CNS, CNS nerve sheath
Brain_Anterior_cingulate_cortex_BA24	CNS, CNS nerve sheath
Brain_Caudate_basal_ganglia	CNS, CNS nerve sheath
Brain_Cerebellar_Hemisphere	CNS, CNS nerve sheath
Brain_Cerebellum	CNS, CNS nerve sheath
Brain_Cortex	CNS, CNS nerve sheath
Brain_Frontal_Cortex_BA9	CNS, CNS nerve sheath
Brain_Hippocampus	CNS, CNS nerve sheath

Brain_Hypothalamus	CNS, CNS nerve sheath
Brain_Nucleus_accumbens_basal_ganglia	CNS, CNS nerve sheath
Brain_Putamen_basal_ganglia	CNS, CNS nerve sheath
Brain_Spinal_cord_cervical_c-1	CNS, CNS nerve sheath
Brain_Substantia_nigra	CNS, CNS nerve sheath
Breast_Mammary_Tissue	Breast
Cells_EBV-transformed_lymphocytes	Haematological lymphoid
Cells_Transformed_fibroblasts	N/A (not site specific)
Colon_Sigmoid	Colorectal
Colon_Transverse	Colorectal
Esophagus_Gastroesophageal_Junction	Oesophagus
Esophagus_Mucosa	Oesophagus
Esophagus_Muscularis	Oesophagus
Heart_Atrial_Appendage	Cardiac myxoma
Heart_Left_Ventricle	N/A (no tumours in series in this tissue)
Liver	N/A (no tumours in series in this tissue)
Lung	Lung
Minor_Salivary_Gland	Salivary gland
Muscle_Skeletal	Soft tissue sarcoma
Nerve_Tibial	PNS nerve sheath benign, PNS nerves heath, Nerve sheath benign
Ovary	Ovary
Pancreas	Pancreas
Pituitary	Pituitary
Prostate	Prostate
Skin_Not_Sun_Exposed_Suprapubic	NMSC, Melanoma, Skin benign
Skin_Sun_Exposed_Lower_leg	NMSC, Melanoma, Skin benign
Small_Intestine_Terminal_Ileum	Small bowel, GINET
Spleen	N/A (no tumours in series in this tissue)
Stomach	Gastric
Testis	Testicular
Thyroid	Thyroid
Uterus	Endometrial, Uterine leiomyoma, Uterine sarcoma
Vagina	N/A (no tumours in series in this tissue)
Whole_Blood	Haematological lymphoid, Haematological myeloid, Haematological polycythaemia, Haematological thrombocythaemia

ACC – Adrenocortical carcinoma, CNS – Central Nervous system, EBV – Epstein Barr Virus, GINET – Gastrointestinal neuroendocrine tumour, NMSC – Non-melanoma skin cancer, PNS – Peripheral nervous system.

The second set of eQTL were reported by a study analysing tumour tissues as opposed to assumed normal tissues from donors.[384] Paired tumour-normal WGS with matched transcriptome was obtained for 930 samples and associations identified between somatic SNVs and expression of target genes proposed by the variant being within a putative regulatory region (as defined by GeneHancer or within 1kb of a transcription start site). eQTL are frequently expressed as regions because SNVs occurring within 50bp of each other are grouped together. Supplementary tables from the publication

resulting from the study were downloaded and higher confidence eQTL (10% false discovery rate cut-off incorporating 102 at 5% cut-off and 67 at 5-10% cut-off) from 22 cancer types were retained for further consideration. Most eQTL were duplicated across cancer types, meaning that 27 unique eQTL were used for downstream analysis (Table 6.9). Given that these variants were identified in cancer tissues, no further selection for eQTL affecting particular genes was performed.

Table 6.9 - Expression quantitative trait loci identified through analysis of cancer tissues

Gene affected	eQTL chromosome	eQTL start	eQTL end	Distance to gene transcription start site (bp)
HYI	1	43824528	43824563	95115
RCSD1	1	167427918	167427936	-171547
LIMS2	2	128439680	128439729	-345
C2orf27A	2	133024749	133024808	544715
C3orf18	3	49823985	49824038	781212
GLYCTK	3	52322011	52322052	196
HERC3	4	88637542	88637550	-876028
TERT	5	1295161	1295253	-45
TIGD6	5	149312169	149312257	67958
C6orf136	6	30704977	30705039	90192
TAS2R5	7	141437957	141437957	-52060
NCALD	8	103118690	103118718	17858
ENPP2	8	120718851	120719000	-67820
PARD3	10	34955724	34955748	148517
TSPAN32	11	2017704	2017713	-305535
TMEM138	11	61735191	61735192	605719
KCNJ5	11	128761332	128761340	23
ACOT1	14	74231057	74231077	227139
EDC3	15	74626537	74626587	361824
HMG20A	15	77965491	77965558	252532
ZNF44	19	13128329	13128457	-722679
ZNF284	19	43772478	43772537	-803790
DHX34	19	47901366	47901512	48901
CA11	19	49660338	49660421	-510929
ZNF551_ZNF544	19	58322231	58322339	128948
SIRPB1	20	1598197	1598223	2479
CTNNBL1	20	36794104	36794104	471747

The resulting genomic coordinates corresponding to both sets of eQTL were used to produce two BED files with which to extract variants from VCFs generated from WGS data as per the truncation variant analysis, although no filter was imposed relating to predicted consequence of the variant.

For eQTL generated by GTEx, a number of phenotypic subgroups of cases (drawn from the same pool as for truncating variant analysis) were subject to case-control analysis according to the tissues in

which eQTL were reported to have an effect. For example, in breast cancer cases, only eQTL altering gene expression in breast tissue would be considered. Initially, all GTEx tissues were designated with tumour labels corresponding to neoplasms occurring in the MPT series that could arise from that tissue (Table A11). For example, adrenal gland tissue was attached to the terms phaeochromocytoma and adrenal cortical carcinoma. 23 phenotypic subgroups of cases were formulated to incorporate all cases with a tumour arising from a GTEx tissue. A group containing all cases was also used (Table 6.10).

Table 6.10 - Phenotypic subgroups used for GTEx expression quantitative trait loci analysis

No. tumours required to be included	Tumours	No. individuals	No. individuals for counts incorporating structural variants
N/A - All MPT individuals	All	424	360
1 From 1	Breast	215	186
1 From 1	Colorectal	98	81
1 From 3	NMSC, Melanoma, Skin benign	78	68
1 From 3	Endometrium, Uterine leiomyoma, Uterine sarcoma	53	44
1 From 1	Ovary	50	42
1 From 4	Haematological lymphoid, Haematological myeloid, Haematological polycythaemia, Haematological thrombocythaemia	40	35
1 From 1	Thyroid	38	27
1 From 1	Haematological lymphoid	33	30
1 From 1	Lung	17	14
1 From 1	Prostate	17	15
1 From 2	CNS, CNS nerve sheath	14	12
1 From 1	Soft tissue sarcoma	13	11
1 From 1	Pituitary	12	9
1 From 2	Small bowel, GINET	11	9
1 From 2	Phaeochromocytoma, ACC	10	9
1 From 1	Pancreas	7	4
1 From 3	PNS nerve sheath benign, PNS nerve sheath, Nerve sheath benign	6	6
1 From 1	Testicular	6	5
1 From 1	Salivary gland	4	4
1 From 1	Oesophagus	3	2
1 From 1	Cardiac myxoma	2	2
1 From 1	Gastric	2	2

ACC – Adrenocortical carcinoma, CNS – Central nervous system, GINET – Gastrointestinal neuroendocrine tumour, GIST – Gastrointestinal stromal tumour, NMSC – Non-melanoma skin cancer, PNS – Peripheral nervous system

Each variant in the variant table produced by filtering was annotated with the corresponding GTEx eQTL identifier, the gene whose expression is affected by it, and the tissue where the association is noted. This annotation was used, for each phenotypic subgroup, to reduce the variant table down to only those eQTL which influence expression in a tissue relevant to that subgroup. Counting of individuals with variants amongst cases vs controls with hypothesis testing was performed as per truncating variant analysis. For counts of individuals with particular variants, the number of tests for correction purposes was taken as the number of unique tissue specific variants in the variant table in cases or controls. For the counts of individuals with variants at eQTL reported to affect the expression of each gene, the number of tests was taken as 83 (number of genes considered that are affected by GTEx eQTL).

As per other analyses described in this chapter, structural variant call data were also interrogated for SVs that affected the considered eQTL. This was performed as per truncating variant analysis but only filtered for deletions of the eQTL (Canvas or Manta calls) because GTEx eQTL are expressed as single nucleotide coordinates and breakpoints are less likely to be relevant. eQTL reported to enhance expression of proto-oncogenes were filtered out because deletion of an eQTL in which variants are associated with downregulation of a tumour suppressor gene is more likely to emulate the effect of an eQTL SNV at that loci than an SV (of any type) is to emulate an eQTL SNV that upregulates expression of a proto-oncogene. Counts of individuals with SVs affecting eQTLs reported to influence expression of each gene were compared between cases and controls using the same phenotypic subgroups, albeit with reduction in numbers due to SV call availability (Table 6.10). Counts of individuals with SVs combined with SNVs/indels were also considered. Subsetting of eQTL according to tissue type was used as for SNVs/indels. The number of tests for Benjamani-Hochberg multiple hypothesis adjustment was again 83.

For the eQTL reported from cancer tissue studies, the broader phenotypic subgroupings used for truncating variant analysis were employed as differences in gene expression contributing to tumourigenesis in one cancer type may be relevant to others. Variant counting within these groups and hypothesis testing was performed as per GTEx based analysis but without any variant sub-setting based on the tissue in which the eQTL was reported. 27 (the number of eQTL considered, each with a unique associated gene) was taken as the number of tests when considering counts of individuals with a variant in an eQTL reported to affect the expression of each gene. SV counts and counts of SVs and SNVs indels affecting these eQTL was also considered in the same way but given that eQTL reported from cancer tissue studies are expressed as regions, deletions (Canvas or Manta) and translocations (Manta) were included as these SV types may be more likely to emulate the effect of variants in those regions than inversions, gains, insertions or duplications.

6.6 - Analysis for causative variants in a family with suspected recessive tumour predisposition

6.6.1 - Introduction

Whilst case control based analyses have identified numerous variants and genes contributing to disease, the basis of multiple genetic conditions has been elucidated by analysing families where multiple members have a phenotype that is considered to be due to the same genetic factor under a hypothesised mode of inheritance. Consideration of the segregation of variants in affected and unaffected family members can reduce the number of putative causative variants, particularly under a recessive hypothesis or in a presumed dominant inheritance pattern where more than one family is available for analysis. The MPT series contained few family members of probands but a single family with a possible autosomal recessive tumour predisposition syndrome was investigated.

Autosomal recessive conditions can be suggested by the occurrence of multiple siblings affected with a similar phenotype that are born to unaffected parents, particularly if both males and females are affected. In the investigated family, a brother and sister had been affected with osteomas and/or lipoma. The female sibling had bilateral mandibular osteomas at age 11 whilst the male sibling had osteoma in an unspecified site (not histologically confirmed) and a 5.5cm (largest dimension) lipoma in the left deltoid region, both before the age of 13 years. There was a further unaffected male sibling aged 9 years and both parents had no history of tumours. There was no family history of neoplasms except for two diagnoses of breast (age 74 years) and prostate (age ~60 years) cancer in the paternal grandmother and a paternal uncle respectively. No consanguinity was reported in the medical record.

The female sibling had had *APC* genetic testing with no deleterious variants identified. The male sibling was identified as harbouring an *ATM* variant that was assessed as likely pathogenic in the WGS-based comprehensive CPG analysis described in Chapter 4. This variant was not present in the sister, however.

WGS had been performed on blood samples from both affected siblings and both parents. Variants resulting from this were analysed according to both a homozygous and compound heterozygous hypothesised mechanism of causation.

6.6.2 - Methods

6.6.2.1 - Variant filtering (Script RA6.23)

Initially, all exonic variants from the four samples were extracted from per chromosome BRIDGE merged VCFs using a pre-prepared hg19 based BED file associated with the Nextera Rapid Capture kit version 1.2 (Illumina Inc., San Diego, CA, USA) and merged into a single VCF. Exonic variants

were chosen because using all variants from WGS would result in a high number of putative causative variants, about which information regarding possible pathogenicity would likely be inadequate for exclusion. The merged VCF was filtered with bcftools filter based on quality parameters (as per truncating variant analysis) to exclude genotypes that didn't meet the specified criteria.

Subsequently, the file was split into per individual VCFs and bcftools view used to output six new per individual files containing variants where the genotype conformed to a specified zygosity. For the affected siblings, files containing only sites with homozygous variants were created as well as files containing only sites with heterozygous variants. Files containing only sites with heterozygous variants were output for both parents.

To check for possible causative variants according to a homozygous hypothesis, bcftools isec was used to output sites that were present in both offspring homozygous VCFs and both parental heterozygous VCFs. The coordinates of any variants fulfilling these criteria where then used to extract variants at those positions from the original merged VCF containing genotype information for all four individuals. The newly subset merged VCF was then annotated with Ensembl VEP and filtered to retain variants with a consequence annotation suggestive of an effect on protein function ("splice_acceptor_variant", "splice_donor_variant", "stop_gained", "frameshift_variant", "stop_lost", "start_lost", "initiator_codon_variant", "inframe_insertion", "inframe_deletion", "missense_variant", "protein_altering_variant") and with an allele frequency in 1000 Genomes European data of <0.05.

Potentially causative variants according to a compound heterozygote hypothesis were generated in a similar manner but bcftools isec was this time used to output separate files containing variant sites present in the heterozygous VCFs from both offspring and one parent, with the process repeated for the other parent. Coordinates generated from both these enquiries were collated and used for a further extraction of variants from the merged VCF.

6.6.2.2 - Review of filtered variants

Variants identified by the above process according to a homozygous hypothesis (n=2) were reviewed further, taking into consideration allele frequencies in publically available datasets[161,166,349] and in the European non-MPT BRIDGE control group as previously utilised in case control analyses. Presence in other MPT cases was also considered. In addition, the GeneCards[380] entry for the relevant genes was reviewed for disease associations and functional descriptions. GeneMania[346] was used to check for interactions between genes containing variants and known CPGs. There are no osteoma studies contained in the cBioportal[34] platform but gene names were entered into it to check for recurrent mutation or expression abnormalities across tumour types.

Variant pairs that were identified by the workflow designed to search for causative compound heterozygous variants were only considered further if they were in the same gene, each parent harboured one of them and both offspring were heterozygous for both variants. Resulting variants were then reviewed in a similar way to those proposed as part the homozygous hypothesis. If either variant in a variant pair had a maximum allele frequency in any 1000 Genomes or gnomAD population above 0.05 then the corresponding variant pair was not considered further.

6.7 - Results

Outputs from the various analyses to detect novel loci potentially involved in cancer predisposition are presented and discussed together in this section. These include case control analysis of truncating variants, variants in putative proto-oncogenes and variants in genes associated with telomere function/maintenance. Also incorporated are analyses of non-coding variants, namely variants within ultra-conserved regions/enhancers/promoters or those within loci associated with altered expression of CPGs reported by either the GTEx project (in normal tissues) or Zhang et al (in cancer tissues).[384]

6.7.1 - Truncating variants in known or suspected cancer predisposition genes (see 6.2)

Counts of individuals harbouring variants in each gene on a gene list were considered and compared with that in a group of controls. Frequency of individual variants was also considered. Analyses were performed utilising multiple gene lists, phenotypic subgroups and zygosity statuses with multiple hypothesis correction applied within each analysis.

Gene level comparisons where the q-value was below a 0.05 significance threshold (n=53) are shown in Table 6.11 whilst comparisons at variant level are described in Table 6.12. These are considered to be the genes/variants most likely to represent causative association with the considered phenotype. Most of the genes/variants have multiple highlighted results, indicating that the result reaches the significance threshold in multiple comparisons using a number of different gene lists, phenotypic subgroups or zygosity states.

Top gene level results were *CHEK2*, *MAX*, *NF1* or *PALB2*, all of which are known CPGs. Significant results involving *CHEK2* were noted in eight phenotypic subgroups, seven of which specifically incorporated breast cancer cases. Individuals with *CHEK2* variants are summarised, along with the variant they harboured, in Table 6.13. Nine participants with c.1229delC (p.Thr410fs) (also referred to as c.1100delC) were recorded as well as five individuals with other variants. Although non-breast tumours were included in most of the subgroups producing significant results, most of the individuals contributing to them (10/11 females) had previously been diagnosed with breast cancer. *CHEK2* truncating variants were also over-represented across all MPT cases (Figure 6.4) due to fourteen

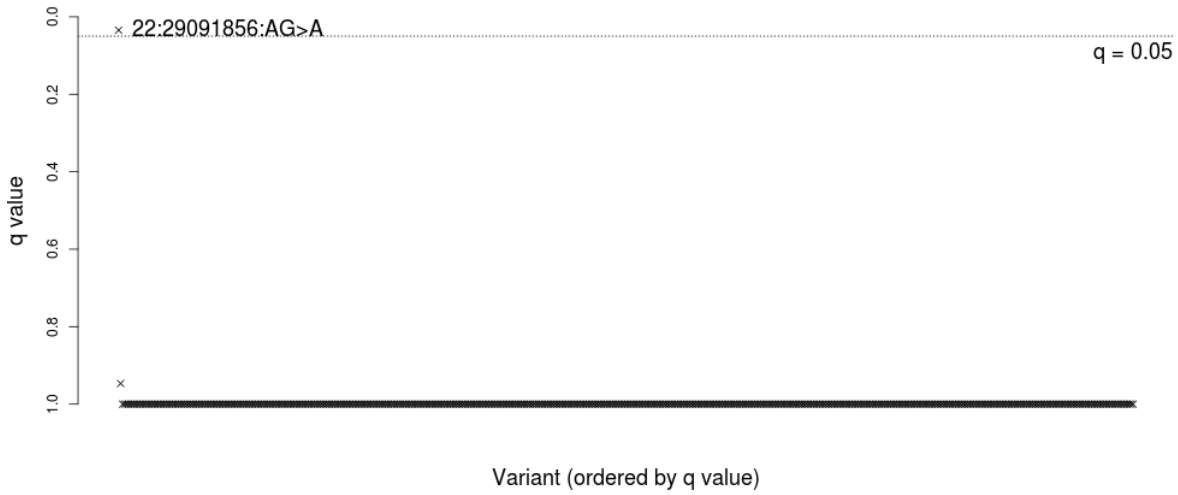
heterozygotes and one homozygote. 8/14 heterozygotes (57.1%, all female) were breast cancer cases whilst 6/14 (42.8%, 5 females and 1 male) had not been diagnosed with that tumour.

Figure 6.4 - Hypothesis tests (individuals with variants per gene) from analysis of all MPT cases (n= 424) - Full gene list (n=1055), heterozygous individuals

Plot shows data points corresponding to gene variants that were present in any BRIDGE or 1958BC sample

The spectrum of non-breast tumours in variant carriers is heterogeneous but the most frequent is renal cell carcinoma (RCC), which occurred in 4/14 (28.6%) heterozygous individuals, three of whom were males who had not developed breast cancer. One individual with RCC was also identified with a translocation affecting *FLCN*. When compared with the 409 MPT individuals without a heterozygous *CHEK2* truncating variant (homozygote for c.1229delC (p.Thr410fs) excluded), the frequency of RCC was not significantly increased at a p-value threshold of <0.05 (4/14 cases in variant carriers vs 52/409 in non-variant carriers, Fishers exact test p = 0.09975). Furthermore, *CHEK2* was not highlighted in the analysis of the RCC phenotypic subgroup (56 individuals). At variant level, eight individuals (1 male and 7 females) with *CHEK2* c.1229delC (p.Thr410fs) led to over-representation (heterozygous or homozygous) in the subgroup diagnosed with breast, thyroid or endometrial cancer (Figure 6.5). Six carriers (all female) had breast cancer, 2 had endometrial cancer and 1 (male) had thyroid malignancy.

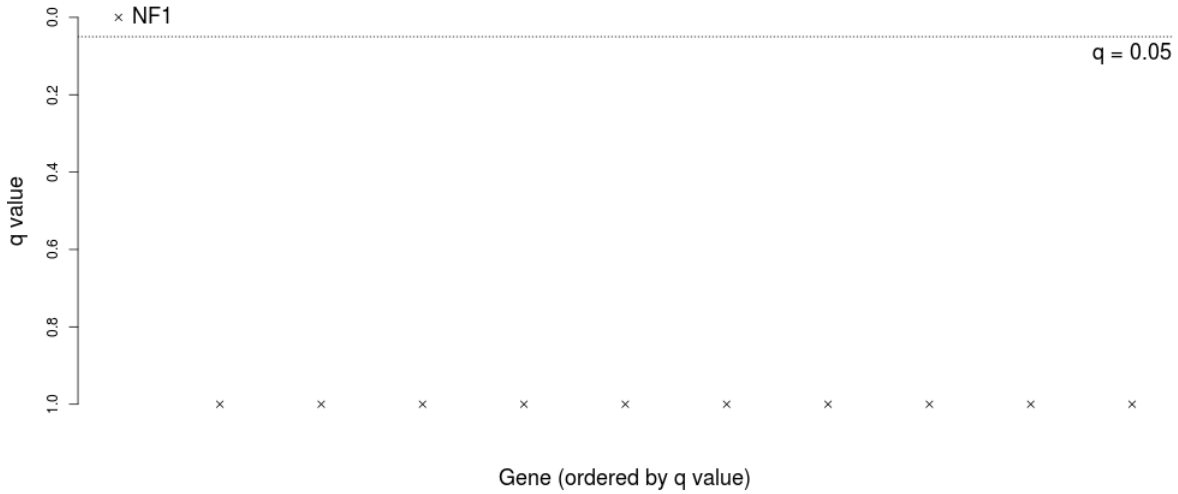
Figure 6.5 - Hypothesis tests (individual variants) from analysis of cases with ≥1 tumour from Breast, thyroid and endometrium (n=260) - Repair gene list (n=445), heterozygous or homozygous individuals. 22:29091856:AG>A corresponds to *CHEK2* c.1229delC (p.Thr410fs)



Plot shows data points corresponding to gene variants that were present in any BRIDGE or 1958BC sample

Truncating variants in *NF1* were over-represented in a number of phenotypic subgroups involving gastrointestinal stromal tumour (GIST) (Figure 6.6), accounted for by four individuals diagnosed with that tumour (Table 6.14). All of these individuals had typical features of Neurofibromatosis type 1 and had previously been diagnosed clinically.

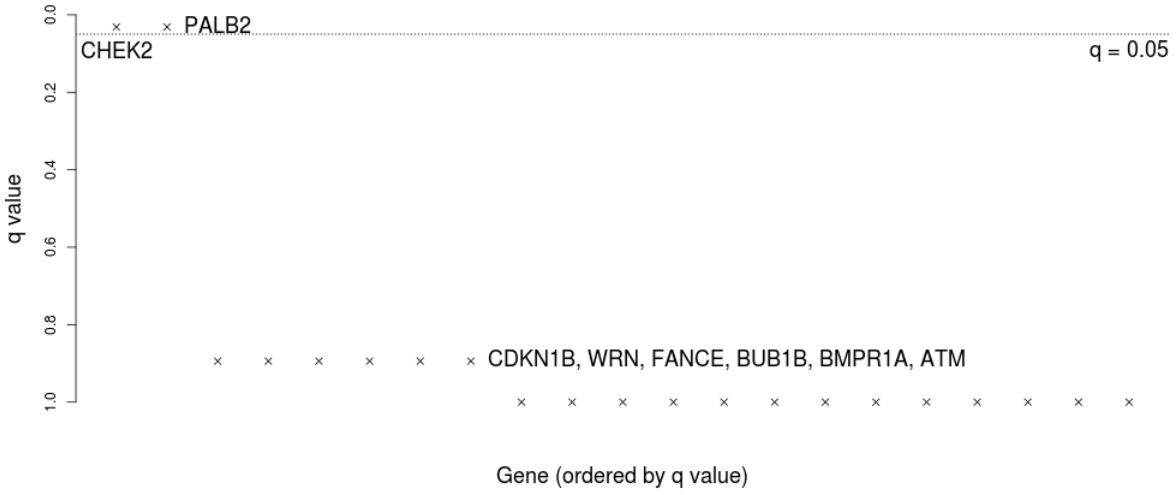
Figure 6.6 - Hypothesis tests (individuals with variants per gene) from analysis of GIST cases (n= 15) - Full gene list (n=1055), heterozygous individuals



Plot shows data points corresponding to gene variants that were present in any BRIDGE or 1958BC sample

PALB2 truncating variants were observed in five breast cancer cases (Table 6.15), leading to individuals with variants in that gene being over-represented in six phenotypic subgroups. These included the group of any individual with breast cancer (n=215, heterozygotes or homozygotes) (Figure 6.7) and other groupings involving breast cancer. There was also over-representation in the subgroup diagnosed with at least one tumour from haematological myeloid, aerodigestive tract, anus or melanoma (tumours associated with *TERT* variants) but all four individuals contributing to this result had also been diagnosed with breast cancer. Three heterozygotes from this subgroup harboured the c.3113G>A (p.Trp1038*) variant, leading to a significantly elevated frequency of this particular variant vs controls (Table 6.15, Figure 6.8)

Figure 6.7 - Hypothesis tests (individuals with variants per gene) from analysis of breast cancer cases (n= 215) - Mania gene list (n=142), heterozygous or homozygous individuals



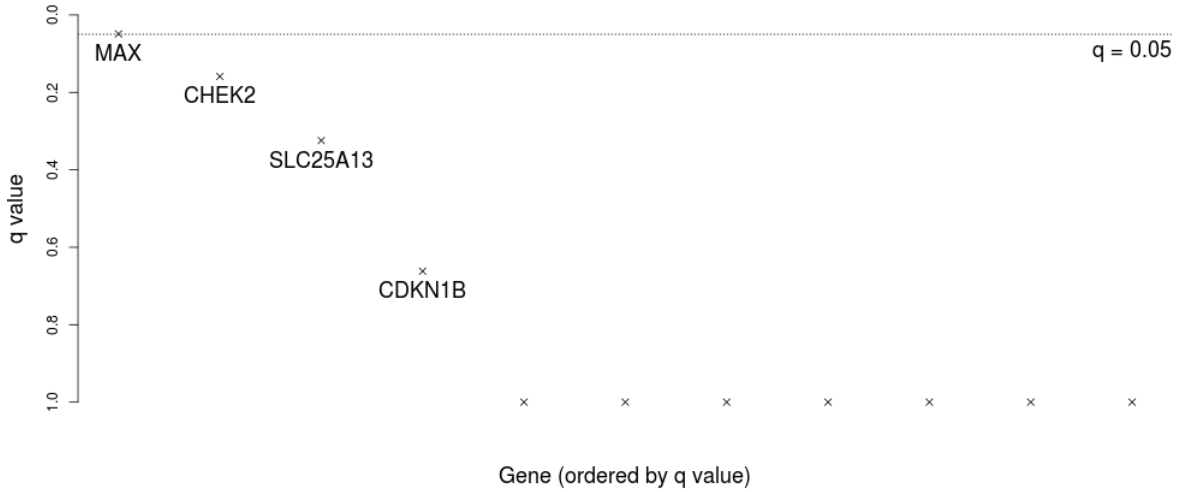
Plot shows data points corresponding to gene variants that were present in any BRIDGE or 1958BC sample

Figure 6.8 - Hypothesis tests (individual variants) from analysis of cases with ≥1 tumour from Haematological myeloid, aerodigestive tract, anus and melanoma (n=52) - Repair gene list (n=445), heterozygous individuals. 16:23632683:C>T corresponds to *PALB2* c.3113G>A (p.Trp1038*)

Plot shows data points corresponding to gene variants that were present in any BRIDGE or 1958BC sample

Two individuals with phaeochromocytoma (Table 6.16) harboured heterozygous truncating variants in *MAX*, leading to q-values below 0.05 when comparing their frequency between controls and individuals with at least one tumour from kidney, phaeochromocytoma, paraganglioma and haemangioblastoma (tumours associated with *VHL* variants) (Figure 6.9).

Figure 6.9 - Hypothesis tests (individuals with variants per gene) from analysis of cases with ≥1 tumour from kidney, phaeochromocytoma, paraganglioma and central nervous system haemangioblastoma (n= 77) - Mania gene list (n=142), heterozygous individuals



Plot shows data points corresponding to gene variants that were present in any BRIDGE or 1958BC sample

Table 6.11 – Genes in which truncating variants over-represented in cases vs controls

Gene	Phenotypic subgroup	Gene list/s	Het cases	Proportion cases het (%)	Het controls	Proportion controls het (%)	q value for hets	Hom cases	Het or hom cases	Proportion cases het or hom (%)	Het or hom controls	Proportion controls het or hom (%)	q value for het or hom
CHEK2	1 From 2 - Breast, Gastric	Mania	7	3	26	0.6	0.074	1	8	4	26	0.6	0.03
CHEK2	1 From 2 - Breast, Ovary	Mania	7	3	26	0.6	0.139	1	8	3	26	0.6	0.05
CHEK2	1 From 2 - Breast, Pancreas	Full	8	4	26	0.6	0.244	1	9	4	26	0.6	0.048
CHEK2	1 From 2 - Breast, Pancreas	Mania	8	4	26	0.6	0.033	1	9	4	26	0.6	0.006
CHEK2	1 From 2 - Breast, Pancreas	Repair	8	4	26	0.6	0.103	1	9	4	26	0.6	0.02
CHEK2	1 From 2 - Breast, Pancreas	Webgestalt	8	4	26	0.6	0.143	1	9	4	26	0.6	0.028
CHEK2	1 From 3 - Breast, Thyroid, Endometrium	Full	9	3	26	0.6	0.17	1	10	4	26	0.6	0.035
CHEK2	1 From 3 - Breast, Thyroid, Endometrium	Mania	9	3	26	0.6	0.023	1	10	4	26	0.6	0.005
CHEK2	1 From 3 - Breast, Thyroid, Endometrium	Repair	9	3	26	0.6	0.072	1	10	4	26	0.6	0.015
CHEK2	1 From 3 - Breast, Thyroid, Endometrium	Webgestalt	9	3	26	0.6	0.099	1	10	4	26	0.6	0.021
CHEK2	1 From 4 - Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal	Mania	8	3	26	0.6	0.085	1	9	3	26	0.6	0.032
CHEK2	1 From 8 - Breast, ACC, CNS, Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma	Mania	7	3	26	0.6	0.092	1	8	3	26	0.6	0.031
CHEK2	All	Full	13	3	26	0.6	0.035	1	14	3	26	0.6	0.009
CHEK2	All	Mania	13	3	26	0.6	0.005	1	14	3	26	0.6	0.001
CHEK2	All	Repair	13	3	26	0.6	0.015	1	14	3	26	0.6	0.004
CHEK2	All	Webgestalt	13	3	26	0.6	0.021	1	14	3	26	0.6	0.005
CHEK2	1 From 1 – Breast	Mania	7	3	26	0.6	0.074	1	8	4	26	0.6	0.03
MAX	1 From 4 - Kidney, Phaeochromocytoma, Paraganglioma, CNS haemangioblastoma	Mania	2	3	0	0	0.049	0	2	3	0	0	0.049
NF1	1 From 3 - Phaeochromocytoma, Paraganglioma, GIST	CGP	4	13	3	0.07	0.00002	0	4	13	3	0.07	0.00002
NF1	1 From 3 - Phaeochromocytoma, Paraganglioma, GIST	Full	4	13	3	0.07	0.0001	0	4	13	3	0.07	0.0001
NF1	1 From 3 - Phaeochromocytoma, Paraganglioma, GIST	Loftool	4	13	3	0.07	0.00005	0	4	13	3	0.07	0.00005
NF1	1 From 3 - Phaeochromocytoma, Paraganglioma, GIST	Mania	4	13	3	0.07	0.00002	0	4	13	3	0.07	0.00002

NF1	1 From 3 - Phaeochromocytoma, Paraganglioma, GIST	Webgestalt	4	13	3	0.07	0.00007	0	4	13	3	0.07	0.00007
NF1	1 From 5 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma	CGP	4	13	3	0.07	0.00002	0	4	13	3	0.07	0.00002
NF1	1 From 5 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma	Full	4	13	3	0.07	0.0001	0	4	13	3	0.07	0.0001
NF1	1 From 5 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma	Loftool	4	13	3	0.07	0.00004	0	4	13	3	0.07	0.00004
NF1	1 From 5 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma	Mania	4	13	3	0.07	0.00001	0	4	13	3	0.07	0.00001
NF1	1 From 5 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma	Webgestalt	4	13	3	0.07	0.00006	0	4	13	3	0.07	0.00006
NF1	1 From 7 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma, Thyroid	CGP	4	4	3	0.07	0.002	0	4	4	3	0.07	0.002
NF1	1 From 7 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma, Thyroid	Full	4	4	3	0.07	0.011	0	4	4	3	0.07	0.011
NF1	1 From 7 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma, Thyroid	Loftool	4	4	3	0.07	0.005	0	4	4	3	0.07	0.005
NF1	1 From 7 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma, Thyroid	Mania	4	4	3	0.07	0.001	0	4	4	3	0.07	0.001
NF1	1 From 7 - Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma, Thyroid	Webgestalt	4	4	3	0.07	0.006	0	4	4	3	0.07	0.006
NF1	1 From 7 - Retinoblastoma, Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma	CGP	4	6	3	0.07	0.0005	0	4	6	3	0.07	0.0005

NF1	1 From 7 - Retinoblastoma, Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma	Full	4	6	3	0.07	0.002	0	4	6	3	0.07	0.002
NF1	1 From 7 - Retinoblastoma, Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma	Loftool	4	6	3	0.07	0.001	0	4	6	3	0.07	0.001
NF1	1 From 7 - Retinoblastoma, Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma	Mania	4	6	3	0.07	0.0003	0	4	6	3	0.07	0.0003
NF1	1 From 7 - Retinoblastoma, Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma, Melanoma	Webgestalt	4	6	3	0.07	0.001	0	4	6	3	0.07	0.001
NF1	1 From 8 - Breast, ACC, CNS, Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma	Mania	4	2	3	0.07	0.042	0	4	2	3	0.07	0.031
NF1	1 From 1 – GIST	CGP	4	27	3	0.07	0.000001	0	4	27	3	0.07	0.000001
NF1	1 From 1 – GIST	Full	4	27	3	0.07	0.000004	0	4	27	3	0.07	0.000004
NF1	1 From 1 – GIST	Loftool	4	27	3	0.07	0.000002	0	4	27	3	0.07	0.000002
NF1	1 From 1 – GIST	Mania	4	27	3	0.07	0.000003	0	4	27	3	0.07	0.000003
NF1	1 From 1 – GIST	Webgestalt	4	27	3	0.07	0.000001	0	4	27	3	0.07	0.000001
PALB2	1 From 2 - Breast, Gastric	Mania	5	2	9	0.2	0.061	0	5	2	9	0.2	0.030
PALB2	1 From 2 - Breast, Ovary	Mania	5	2	9	0.2	0.01	0	5	2	9	0.2	0.05
PALB2	1 From 2 - Breast, Pancreas	Mania	5	2	9	0.2	0.033	0	5	2	9	0.2	0.033
PALB2	1 From 4 – Haematological myeloid, Aerodigestive tract, Anus, Melanoma	Full	4	8	9	0.2	0.016	0	4	8	9	0.2	0.016
PALB2	1 From 4 – Haematological myeloid, Aerodigestive tract, Anus, Melanoma	Mania	4	8	9	0.2	0.002	0	4	8	9	0.2	0.002
PALB2	1 From 4 – Haematological myeloid, Aerodigestive tract, Anus, Melanoma	Repair	4	8	9	0.2	0.007	0	4	8	9	0.2	0.007

PALB2	1 From 4 – Haematological myeloid, Aerodigestive tract, Anus, Melanoma	Webgestalt	4	8	9	0.2	0.009	0	4	8	9	0.2	0.009
PALB2	1 From 8 - Breast, ACC, CNS, Connective tissue soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma	Mania	5	2	9	0.2	0.05	0	5	2	9	0.2	0.033
PALB2	1 From 1 – Breast	Mania	5	2	9	0.2	0.06	0	5	2	9	0.2	0.03

ACC – Adrenocortical carcinoma, CNS – Central nervous system, GIST – Gastrointestinal stromal tumour, Het – Heterozygous, Hom - Homozygous

Table 6.12 – Truncating variants over-represented in cases vs controls

Gene	Transcript	Coordinate	Description	Gene list	Phenotypic subgroup	Het cases	Het controls	q value for hets	Hom cases	Hom controls	Het or hom cases	Het or hom controls	q value for hets or homs
PALB2	ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Repair	1 From 4 – Haematological myeloid, Aerodigestive tract, Anus, Melanoma	3	3	0.019	0	0	3	3	0.019
PALB2	ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Mania	1 From 4 – Haematological myeloid, Aerodigestive tract, Anus, Melanoma	3	3	0.024	0	0	3	3	0.024
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Repair	1 From 3 – Breast, Thyroid, Endometrium	7	17	0.19	1	0	8	17	0.034
CHEK2	ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Mania	1 From 3 – Breast, Thyroid, Endometrium	7	17	0.25	1	0	8	17	0.045

Table 6.13 – Truncating variants in *CHEK2* (heterozygous)

Transcript	Coordinate	Description	Consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
ENST00000382580	chr22:29091226	c.1392delT (p.Leu464fs)	Frameshift	Kidney, 56; Kidney, 56; Kidney, 56	Unavailable
ENST00000382580	chr22:29091226	c.1392delT (p.Leu464fs)	Frameshift	Kidney, 56; Kidney, 60	Mother - Breast, 47
ENST00000382580	chr22:29091226	c.1392delT (p.Leu464fs)	Frameshift	Thymus, 53; Breast, 54; Haematological lymphoid, 63; Haematological lymphoid, 67	Daughter - Ovary, 41, Colorectal, 41; Paternal uncle - Lung, 76; Paternal uncle - Lung, 78
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)*	Frameshift	Fibrofolliculoma (multiple), 18; Kidney (clear cell carcinoma), 53	Unavailable
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 52; Melanoma, 54	Mother - Breast, <45; Sister - NMSC, <58; Maternal aunt - Ovary, >59
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Endometrium, 54; Breast, 57	Brother - Colorectal, 28; Maternal aunt x2 - Unknown primary ? Age.
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 50 (DCIS); Kidney, 62; GINET (appendix neuroendocrine tumour), 63; Haematological myeloid (CML), 65	Daughter - Neuroendocrine tumour of appendix, 25; Maternal aunt - CLL, 63; Maternal aunt - Breast, 50
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 31; Gastric, 49	Nil
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 45; Breast, 54; Breast (DCIS), 55	Maternal aunt - Gastric, 65
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Colorectal (ascending colon), 27; Endometrium, 53; Colorectal (hepatic flexure), 56; NMSC (multiple BCC), <64	Father - Liver ? age; Paternal uncle - Colorectal ? age.
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Thyroid, 45; Pancreas, 48	Unavailable
ENST00000382580	chr22:29091856	c.1229delC (p.Thr410fs)	Frameshift	Breast, 40; Pancreas benign (solid pseudopapillary tumour), 41	Paternal lineage: Father - Parotid, ? age; Paternal aunt - Breast, 42; Paternal grandmother - Kidney 80. Maternal lineage: Mother - Breast, 54; Maternal cousin - Breast, 39; Maternal aunt - Lung, 53.
ENST00000382580	chr22:29105993	c.1051+1C>T	Splice site (donor)	Breast, 46; Ovary, 49 (bilateral endometrioid); Ovary (bilateral endometrioid), 49; Endometrium (endometrioid), 49	Sister - Breast, 49; Mother - Breast, 46; Maternal uncle - Bladder, 50-59; Maternal grandmother - Breast, 50-59
ENST00000382580	chr22:29115410	c.784delG (p.Glu262fs)	Frameshift	Colorectal polyps (tubulovillous adenomas), 46; Parathyroid (adenoma), 48; Parathyroid (adenoma), 55; Parathyroid (adenoma), 59	Mother - Lung, 53

*Also has translocation with breakpoint in FLCN. BCC – Basal cell carcinoma, CLL – Chronic lymphocytic leukaemia, CML – Chronic Myeloid Leukaemia, DCIS – Ductal carcinoma in-situ, GINET – Gastrointestinal stromal tumour, NMSC – Non-melanoma skin cancer

Table 6.14 – Truncating variants in *NF1* (heterozygous)

Transcript	Coordinate	Description	Consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
ENST00000358273	chr17:29546035	c.1541-1542delAG (p.Gln514fs)	Frameshift	Nerve sheath benign (multiple cutaneous neurofibromas), <30; GIST (jejunal), 46; CNS Nerve sheath (spinal neurofibroma), 51	Clinically NF1 in proband, brother, mother and further 2nd degree relatives. Brother - Adrenal gland tumour, ? age; Mother - Bone tumour ? age.
ENST00000358273	chr17:29684007	c.7768-7769insA (p.His2590fs)	Frameshift	PNS Nerve sheath (MPSNT), 20; GIST (wild type duodenal), 41	Father - Phaeochromocytoma, >59.
ENST00000358273	chr17:29588770	c.4620delA (p.Ala1540fs)	Frameshift	Lipoma (back), 29; GIST (duodenal), 44	Clinically NF1 in proband, mother and maternal grandfather. Mother - Unknown primary, 40
ENST00000358273	chr17:29661873	c.5831delT (p.Leu1944fs)	Frameshift	GIST (multiple jejunal), 36	Clinically NF1 in proband, daughter, brother and mother. Brother - Optic glioma, 5; Daughter - Rhabdomyosarcoma, 3

CNS – Central nervous system, NF1 - Neurofibromatosis type 1, GIST – Gastrointestinal stromal tumour, MPSNT – Malignant peripheral nerve sheath tumour, PNS – Peripheral nervous system.

Table 6.15 - Truncating variants in *PALB2* (heterozygous)

Transcript	Coordinate	Description	Consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
ENST00000261584	chr16:23625409	c.3116delA (p.Asn1039fs)	Frameshift	Breast, 35; Skin sarcoma (angiosarcoma in radiotherapy field), 37; Aerodigestive tract (nasal cavity SCC), 43	Sister - Breast, 48; Mother - Breast, 35
ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Stop gain	Anus, 37; Breast, 42	Father - Gastric, 69; Paternal uncle - NMSC, 66; Paternal grandmother - Unknown primary, 87; Paternal cousin once removed - Unknown primary, 48; Paternal cousin once removed - Aerodigestive tract, 48; Paternal great aunt - Ovary, 53; Paternal great uncle - Liver, 40; Paternal great uncle - Lung, 60; Paternal great grandfather - Gastric, 43
ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Stop gain	Melanoma, 31; Breast, 40	Father - Breast, 68; Paternal great aunt - Breast, 30
ENST00000261584	chr16:23632683	c.3113G>A (p.Trp1038*)	Stop gain	Melanoma, 38; Breast, 47	Sister - Breast, 58; Sister - Breast, 51; Mother - Breast, 48; Maternal grandmother - Breast, 48; Maternal cousin - Breast, 43
ENST00000261584	chr16:23649437	c.62T>G (p.Leu21*)	Stop gain	Colorectal, 51; Breast, 54	Sister - Breast, 43; Sister - Breast, 43; Sister - NHL, 53; Brother - Prostate, 67; Brother - Colorectal, 40; Paternal grandfather - Colorectal, 65

NHL – Non-Hodgkin's lymphoma, NMSC – Non-melanoma skin cancer, SCC – Squamous cell carcinoma

Table 6.16 – Truncating variants in *MAX* (heterozygous)

Transcript	Coordinate	Description	Consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
ENST00000358664	chr14:65544637	c.289C>T (p.Gln97*)	Stop gain	Phaeochromocytoma, 16; Phaeochromocytoma, 35	Sister - Phaeochromocytoma, <49
ENST00000358664	chr14:65560500	c.228C>T (p.Arg33*)	Stop gain	Phaeochromocytoma, 43; Kidney, 43	Father - Testicular, 60-69

Counts of structural variants affecting each gene on the gene lists were also compared in cases vs controls. One result returned a q-value below the chosen significance threshold of 0.05 and was produced by the occurrence of a heterozygous translocation (review of BAM file with Integrated Genomics Viewer (IGV) showed some reads supporting this call but only when viewed from one end of the translocation (Appendix 5, variant 8)) affecting *HABP2* in a single individual with breast, colorectal and pancreatic cancer (Table 6.17) vs two controls in nine phenotypic subgroups (All MPT cases; 1 From 3 Breast, Thyroid, Endometrium; 1 From 8 Breast, Adrenocortical carcinoma (ACC), Central nervous system (CNS) tumours, Soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma; 1 From 4 Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal; 1 From 2 Breast, Gastric; 1 From 1 Breast; 1 From 2 Breast, Pancreas; 1 From 4 Breast, Aerodigestive tract, Lung , Ovary and 1 From 2 Breast Ovary).

This individual with the *HABP2* translocation also contributed to one of a number of gene level results with q-values below the significance thresholds when counts of individuals with SVs and SNVs/indels were combined (Table 6.18). Three individuals with heterozygous *HABP2* truncating variants contributed to a total of three cases among two subgroups producing such results (Tables 6.19 and 6.20). One individual with a heterozygous nonsense *BMPR1A* variant and a one with a translocation affecting that gene produced a significant result in the analysis of all MPT cases as well as other phenotypic subgroups involving breast cancer (two cases vs 11 controls) (Tables 6.21 and 6.22). The *BMPR1A* translocation was predicted to have a breakpoint between exons 1 and 2, which are both non-coding. Review of reads supporting the variant call in IGV demonstrated that all of them were due to discordant mate pairs (rather than split reads) aligning to chromosome 5, where the counterpart predicted breakpoint was located, and chromosome 10, where *BMPR1A* is located (Appendix 5, variant 9). Other highlighted results were produced due to a single case with an SNV or indel in *APCS* or *MSH6*. These latter results were not considered further as no contribution to them was made by the addition of SVs to the analysis. The reduction in q-value is likely to have been due to the reduction in size of phenotypic subgroups (due to non-availability of SV calls for some individuals) leading to individuals with variants making up a greater proportion of cases.

Table 6.17 - Predicted structural variant affecting *HABP2* (heterozygous)

Chromosome	Predicted start	Predicted end	Algorithm	Predicted consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
10	Chr10:115318616	chr6:7227789	Manta	Translocation with breakpoint between exons 1 and 2 (both coding)	Breast (bilateral), 46; Colorectal, 51; Pancreas, 52	Mother - Fibrosarcoma, 50; Maternal aunt - Breast, 70; Maternal grandfather – Prostate, 52

Table 6.18 – Genes in which truncating variants over-represented in cases vs controls where combination of counts of single nucleotide variants, indels and structural variants considered

Gene	Phenotypic subgroup	Gene list/s	Het cases	Proportion cases het (%)	Het controls	Proportion controls het (%)	q value for hets
HABP2	1 From 4 - Colorectal, Breast, Gastric, Ovarysexcord-gonadalstromal	full	3	1.30	45	1.16	0.001
HABP2	1 From 8 - Breast ACC CNS Connective tissue soft tissue sarcoma Bonesarcoma GIST Skinsarcoma Uterinesarcoma	full	3	1.43	45	1.16	0.001
BMPR1A	1 From 2 - Breast Gastric	mania	2	1.08	11	0.28	0.042
BMPR1A	1 From 2 - Breast Pancreas	mania	2	1.06	11	0.28	0.02
BMPR1A	1 From 3 - Breast Thyroid Endometrium	mania	2	0.91	11	0.28	0.013
BMPR1A	1 From 4 - Colorectal Breast Gastric Ovarysexcord-gonadalstromal	mania	2	0.87	11	0.28	0.042
BMPR1A	All	mania	2	0.56	11	0.28	0.005
BMPR1A	1 From 1 – Breast	mania	2	1.08	11	0.28	0.042
MSH6	1 From 3 - Haemmyeloid Aerodigestivetract Anus	mania	1	7.69	10	0.26	0.021
MSH6	1 From 4 - Haemmyeloid Aerodigestivetract Anus Melanoma	mania	1	2.22	10	0.26	0.003
APCS	1 From 4 - Colorectal Breast Gastric Ovarysexcord-gonadalstromal	full	1	0.43	10	0.26	0.047
APCS	1 From 4 - Colorectal Endometrium Ovary Sebaceous	full	1	0.70	10	0.26	0.025
APCS	1 From 8 - Breast ACC CNS Connectivetissuesofttissuesarcoma Bonesarcoma GIST Skinsarcoma Uterinesarcoma	full	1	0.48	10	0.26	0.04
APCS	All	full	1	0.28	10	0.26	0.0001

ACC – Adrenocortical carcinoma, CNS – Central nervous system, GIST – Gastrointestinal stromal tumour, Het - Heterozygous, Hom – homozygous

Table 6.19 - Truncating variants in *HABP2* (heterozygous) amongst 1 From 4 Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal (Peutz-Jeghers like) phenotypic subgroup

Transcript	Coordinate	Description	Consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
ENST00000351270	chr10: 115341778	c.982C>T (p.Q328*)	Stop gain	Thrombocythaemia, 56; Breast, 56	Mother – Breast, 79; Daughter – Breast, 34; Maternal aunt – Breast, 80
ENST00000351270703	chr10: 115338424	c.607C>T (p.R203*)	Stop gain	Retinoblastoma, 2; Colorectal, 49	Father – Colorectal, 79; Paternal cousin – Colorectal, 55

Table 6.20 - Truncating variants in *HABP2* (heterozygous) amongst 1 From 8 Breast, ACC, CNS, Soft tissue sarcoma, Bone sarcoma, GIST, Skin sarcoma, Uterine sarcoma (Li Fraumeni like) phenotypic subgroup

Transcript	Coordinate	Description	Consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
ENST00000351270	Chr10:115338424	c.607C>T (p.R203*)	Stop gain	GIST, 16; Paraganglioma, <30	Nil
ENST00000351270703	Chr10:115341778	c.982C>T (p.Q328*)	Stop gain	Thrombocythaemia, 56; Breast, 56	Mother – Breast, 79; Daughter – Breast, 34; Maternal aunt – Breast, 80

GIST – Gastrointestinal stromal tumour

Table 6.21 - Predicted structural variant affecting *BMPR1A* (heterozygous)

Chromosome	Predicted start	Predicted end	Algorithm	Predicted consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
10	Chr10:88559247	chr5:107163219	Manta	Translocation with breakpoint between exons 1 and 2 (both non-coding)	Breast, 52; CNS meningioma, 56; Breast, 58; Aerodigestive tract, 63	Paternal grandmother – Unknown primary, 75

CNS – Central nervous system

Table 6.22 - Truncating variant in *BMPR1A* (heterozygous)

Transcript	Coordinate	Description	Consequence	Phenotype with age at diagnosis	Family history of neoplasia reported
ENST00000372037	chr10:88676945	c.730 C>T (p.R244*)	Stop gain	Colorectal, 50; Breast, 57	Paternal aunt – Ovarian, 70-79

6.7.2 - Enhancers and promoters (see methods in 6.5.2.1)

Analysis of variants affecting enhancers and promoters of CPGs yielded no results with q-values below the chosen significance threshold when counts of individuals SNVs/indels were compared or when SVs were considered in a separate analysis.

When counts of individuals with either an SNV/indel or SV affecting each enhancer/promoter were considered, one result was highlighted in the 2 From 3 Breast, Thyroid, Endometrium subgroup. Here, five cases (14%) had heterozygous SNVs or indels affecting enhancer GH17G058351 vs 299 (8%) controls (q=0.02). GH17G058351 is reported to be a *RAD51C* enhancer but ovarian cancer (associated with *RAD51C* variants) is not part of this phenotypic subgroup and no SVs accounted for the five cases. Therefore, the reduction in q-value compared to other analyses is likely due to the reduction in phenotypic subgroup size from 43 to 35, leading to individuals with variants representing a greater proportion of the subgroup.

6.7.3 - Expression quantitative trait loci observed in cancer tissues (see methods in 6.5.2.3)

Case control analyses comparing frequency of variants in eQTL observed in cancer tissues was made as per truncating variants. Given that these are non-coding regions, counts of individuals with variants within a given gene were replaced with counts of individuals harbouring a variant within an eQTL reported to affect the expression of a gene.

Individuals with variants reported to affect the expression of three genes (*TAS2R5*, *ENPP2* and *C2orf27A*) contributed to observed results with a q-value below 0.05 (Table 6.23).

The occurrence of the chr7:141437957 T>C variant (reported to affect *TAS2R5* expression) in four individuals accounted for significant results at both gene (Table 6.23) and variant (Table 6.24) level in a number of phenotypic subgroups, all including colorectal or aerodigestive tract cancer (Figure 6.10). Individuals with this variant are described in Table 6.25. chr7:141437957 T>C is reported to reduce *TAS2R5* expression in breast invasive carcinoma, colon/rectum adenocarcinoma, acute myeloid leukaemia, kidney chromophobe tumours, kidney renal papillary cell carcinoma, glioblastoma multiformae, lung adenocarcinoma and sarcoma. Between two and four individuals harboured the variant in each subgroup. Two individuals had been identified in an earlier analysis (see Chapter 4) as harbouring homozygous pathogenic *NTHL1* variants. No sequencing quality issues were evident with the variants on review of bam files with IGV.

Figure 6.10 - Hypothesis tests (individual variants in cancer tissue eQTL) from analysis of colorectal cases (n=98) - Heterozygous individuals.

Plot shows data points corresponding to variants that were present in any BRIDGE or 1958BC sample

Frequency of heterozygous variants in eQTL upregulating *ENPP2* expression in acute myeloid leukaemia, colon/rectum adenocarcinoma, lung squamous cell carcinoma and ovarian serous cystadenocarcinoma was found to be significantly elevated in individuals diagnosed with both breast and ovarian cancer. Ten individuals with variants (Tables 6.26 and 6.27) in that eQTL region were recorded but one of these variants was excluded following review of the relevant bam file in IGV. The count of individuals with variants was therefore more likely to be 9/24 (37.5%) vs 583/4053 (14%) controls.

2/14 (14%) individuals with both breast and kidney cancer had variants (Table 6.28) in an eQTL associated with reduced expression of *C2orf27A* in glioblastoma multiformae, lung adenocarcinoma, ovarian serous cystadenocarcinoma, prostate adenocarcinoma, sarcoma and stomach adenocarcinoma compared with 13/4053 (0.003%) controls, a sufficient count to produce a q value of <0.05. Review of BAM files with IGV showed multiple reads supporting the variant call but the relevant bases were predominantly covered by reads with low mapping quality.

Table 6.23 - Genes where variants at expression quantitative trait loci reported to affect expression are over-represented in cases vs controls

Gene	Phenotypic subgroup	Zygosity considered	Case count	Proportion cases (%)	Controls count	Proportion controls (%)	q value
TAS2R5	1 From 1 - Aerodigestive tract	Het	2	18	9	0.2	0.01
TAS2R5	1 From 1 - Aerodigestive tract	Het or hom	2	18	9	0.2	0.01
ENPP2	2 From 2 – Breast, Ovary	Het or hom	11	46	656	16	0.018
TAS2R5	1 From 3 - Haematological myeloid, Aerodigestive tract, Anus	Het	2	12	9	0.2	0.024
TAS2R5	1 From 3 - Haematological myeloid, Aerodigestive tract, Anus	Het or hom	2	12	9	0.22	0.024
ENPP2	2 From 2 – Breast, Ovary	Het	10	42	583	14	0.03
C2orf27A	2 From 2 – Breast, Kidney	Het	2	14	13	0.3	0.03
C2orf27A	2 From 2 – Breast, Kidney	Het or hom	2	14	13	0.3	0.03

Het – Heterozygous, Hom - Homozygous

Table 6.24 – Variants in somatic expression quantitative trait locus (region 1bp in length) where variants reported to reduce *TAS2R5* expression

Coordinate	Ref	Alt	Phenotypic subgroup	Coefficient	Distance to transcription start site (bp)	Case count	Controls count (n =4053)	Zygosity considered	q value	Participant with variant (see Table 6.25)
Chr7:141437957	T	C	1 From 1 - Aerodigestive tract	-0.57	-52060	2	1	Het	0.01	1,2
Chr7:141437957	T	C	1 From 1 - Aerodigestive tract	-0.57	-52060	2	1	Het or hom	0.01	1,2
Chr7:141437957	T	C	1 From 3 - Haematological myeloid, Aerodigestive tract, Anus	-0.57	-52060	2	1	Het	0.025	1,2
Chr7:141437957	T	C	1 From 3 - Haematological myeloid, Aerodigestive tract, Anus	-0.57	-52060	2	1	Het or hom	0.025	1,2
Chr7:141437957	T	C	1 From 1 – Colorectal	-0.57	-52060	3	1	Het	0.026	1,2,3
Chr7:141437957	T	C	1 From 1 – Colorectal	-0.57	-52060	3	1	Het or hom	0.026	1,2,3
Chr7:141437957	T	C	1 From 2 – Colorectal, Gastric	-0.57	-52060	3	1	Het	0.027	1,2,3
Chr7:141437957	T	C	1 From 2 – Colorectal, Gastric	-0.57	-52060	3	1	Het or hom	0.02689	1,2,3
Chr7:141437957	T	C	1 From 4 – Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal	-0.57	-52060	4	1	Het	0.038	1,2,3
Chr7:141437957	T	C	1 From 4 – Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal	-0.57	-52060	4	1	Het or hom	0.038	1,2,3,4

Coefficient describes effect and magnitude of effect on gene expression of variants in eQTL (range -1.29 – 1.15 amongst higher confidence eQTL reported in cancer tissues)
Het – Heterozygous, Hom - Homozygous

Table 6.25 – Four cases with chr7:141437957 T>C variant (heterozygous) contributing to statistically significant results involving eQTL where variants reported to reduce *TAS2R5* expression.

Participant	Phenotype with age at diagnosis	Family history of neoplasia reported	Clinically relevant coding variants detected
1	Aerodigestive tract, 50; Breast, 66; Lung, 67; Colorectal, 68	Nil	Nil
2	Colorectal, 48; Aerodigestive tract, 50	Sister – Colorectal, 57, Breast, 51; Sister – Colorectal, 44; Brother – Pancreas, 50; Nephew – Lymphoma, 24; Maternal uncle – Prostate, 55; Maternal grandfather – Leukaemia, 50	*NTHL1* ENST00000219066 c.268C>T (p.Gln90*) homozygote. Assessed as pathogenic (see Chapter 4)
3	CNS meningioma, 42; CNS meningioma, 42; Colorectal, 44	Unavailable	*NTHL1* ENST00000219066 c.268C>T (p.Gln90*) homozygote. Assessed as pathogenic (see Chapter 4)
4	NMSC, 40; Ovary sex cord-gonadal stromal (sertoli leydig), 55	Sister – Breast, 45	Nil

CNS – Central nervous system, NMSC – Non-melanoma skin cancer

Table 6.26– Variants in eQTL where variants reported to increase *ENPP2* expression (heterozygous)

Coordinate	Reference allele	Alternate allele	eQTL start	eQTL end	Participant with variant (see Table 6.27)	Coefficient	Case count (n=24)
chr8:120719000	TTTTC	T	120718851	120719000	1	1.15	1
chr8:120718978	T	TTTTC	120718851	120719000	2, 3	1.15	2
chr8:120718978	T	TTTTCTTTCTTTCTTTCTTTC	120718851	120719000	4, 5, 6	1.15	3
chr8:120718999	T	C	120718851	120719000	7	1.15	1
chr8:120718978	T	TTTTCTTTCTTTCTTTCTTTCTTTCTTTC	120718851	120719000	5	1.15	1
chr8:120718980	TTC	T	120718851	120719000	8	1.15	1
chr8:120718864	CT	C	120718851	120719000	9	1.15	1
chr8:120718982	CTTTCTTTCTTTCTTTCTT	C	120718851	120719000	10	1.15	1

Coefficient describes effect and magnitude of effect on gene expression of variants in eQTL (range -1.29 – 1.15 amongst higher confidence eQTL reported in cancer tissues)

Table 6.27 – Summary of cases with variants in eQTL where variants reported to affect *ENPP2* expression

Participant	Phenotype with age at diagnosis	Family history of neoplasia reported	Clinically relevant coding variants detected
1	Ovary, 34; Breast, 47	Mother – NMSC, 69; Maternal grandfather – Colorectal, 54	Nil
2	Breast, 42; Ovary, 47	Mother – Breast, 56; Maternal uncle – Myeloma, ? age	Nil
3	Breast, 27; Ovary, 49; Endometrium, 49	Paternal grandmother – Unknown primary, 67	Nil
4 (no evidence of variant on review with IGV)	Breast, 46; Ovary, 49	Father – Colorectal, 44; Paternal uncle – Lung, ? age; Paternal cousin – Breast, ? age; Paternal cousin – Breast, ? age; Paternal cousin – Unknown primary, ? age	Nil
5	Ovary, 60; Endometrium, 60; Breast, 62	Sister (monozygotic twin) – Breast, 58; Maternal aunt – Gastric, ? age	Nil
6	Breast, 46; Ovary, 49; Ovary, 49; Endometrium, 49.	Sister – Breast, 49; Mother – Breast, 46; Maternal grandmother – Breast, 50-59; Maternal uncle – Bladder, 50-59	*CHEK2* splice donor variant
7	Ovary, 49; Breast, 50	Maternal grandmother – Gastric, 55	Nil
8	Breast, 48; Ovary, 53; Endometrium, 53; Cervix, 53	Sister – Breast, 63; Niece – Breast, 48; Maternal aunt – Colorectal, ? age; Maternal uncle – Colorectal, ? age	Nil
9	Breast, 60; Breast, 65; Ovary, 67	Nil	Nil
10	Breast, 54; Breast, 54; Oesophagus, 54; Ovary, 67	Daughter – Colorectal, 34; Mother – NMSC, 87; Maternal grandmother – Breast, 42; Sister (half) – Breast, 60-69	Nil

IGV – Integrated Genomics Viewer, NMSC – Non-melanoma skin cancer

Table 6.28 – Single nucleotide variant and indel in eQTL region where variants reported to reduce *C2orf27A* expression (heterozygous)

Coordinate	Reference allele	Alternate allele	eQTL start	eQTL end	Coefficient	Distance to transcription start site (bp)	Case count (n=14)	Phenotype with age at diagnosis	Family history	Clinically relevant coding variants detected
chr2:133024749	G	C	133024749	133024808	-1.23	544715	1	Haematological lymphoid (NHL), 57; Breast, 64; Kidney (papillary type 2), 65; Colorectal, 72; Colorectal, 72	Mother – Breast, 42; Sister – Kidney, 49	Nil
chr2:133024753	CAG	C	133024749	133024808	-1.23	544715	1	Thyroid, 32; Kidney (? subtype), 58; Breast, 63	Unavailable	Nil

Coefficient describes effect and magnitude of effect on gene expression of variants in eQTL (range -1.29 – 1.15 amongst higher confidence eQTL reported in cancer tissues
NHL – Non-Hodgkin's lymphoma

No results with q-values below the chosen significance threshold of 0.05 were produced by analysis of SVs affecting eQTL observed in cancer tissues. When counts of individuals with SVs per eQTL were combined with counts of individuals with an SNV or indel per eQTL, a number of results with q-values <0.05 were produced (Table 6.29) but none of these except one were contributed to by SV counts and are likely to be due to a reduction in the phenotypic subgroup size used, leading to cases with variants representing a greater proportion of the subgroup compared with analyses only involving SNVs and indels.

The result that was contributed to by both SVs and SNVs/indels was the over-representation of individuals with heterozygous variants affecting eQTL reported to affect *ZNF284* expression amongst cases with at least one tumour from colorectal, breast, gastric or ovary sex cord-gonadal stromal (*STK11* like) vs controls (Table 6.30 and 6.31). One individual with a predicted deletion of the eQTL (Appendix 5, variant 10) and one individual with an SNV within it contributed to the result but no tumour types were common to both of them. Additionally, it was difficult to find evidence to support or refute the presence of the large predicted deletion through review of the BAM file with IGV.

Table 6.29 - Genes where eQTL affecting expression over-represented in cases vs controls where combination of counts of single nucleotide variants, indels and structural variants considered

Gene	Phenotypic subgroup	Het cases	Proportion cases het (%)	Het controls	Proportion controls het (%)	q value for hets	Hom cases	Hom controls	Het or hom cases	Proportion cases het or hom (%)	Het or hom controls	Proportion controls het or hom (%)	q value for het or hom
C6orf136	1 From 2 - Breast, Ovary	65	31.55	816	20.98	0.044	1	17	66	32.04	833	21.42	0.049
C6orf136	1 From 4 - Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal	71	30.74	816	20.98	0.043	2	17	73	31.60	833	21.42	0.043
C6orf136	2 From 2 - Breast, Kidney	5	38.46	816	20.98	0.021	1	17	6	46.15	833	21.42	0.021
HERC3	1 From 2 - Colorectal, Gastric	3	3.66	71	1.83	0.044	0	0	3	3.66	71	1.83	0.044
HERC3	1From 4 - Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal	8	3.46	71	1.83	0.031	0	0	8	3.46	71	1.83	0.031
HERC3	1 From 1 – Colorectal	3	3.70	71	1.83	0.043	0	0	3	3.70	71	1.83	0.043
ZNF284	1 From 2 - Breast, Ovary	1	0.49	5	0.13	0.044	0	0	1	0.49	5	0.13	0.050
ZNF284	1 From 4 - Colorectal, Breast, Gastric, Ovary sex cord-gonadal stromal	2	0.87	5	0.13	0.031	0	0	2	0.87	5	0.13	0.031

Het - Heterozygous, Hom - homozygous

Table 6.30 - Structural variant affecting eQTL where variants reported to reduce *ZNF284* expression (heterozygous)

Chromosome	Predicted start	Predicted end	Algorithm	Predicted consequence	eQTL start	eQTL end	Coefficient	Distance to transcription start site (bp)	Phenotype with age at diagnosis	Family history of neoplasia reported	Clinically relevant coding variants detected
19	43765327	43848192	Canvas	Deletion of entire eQTL region	43772478	43772537	-0.99	-803790	Prostate, 54; Colorectal, 54	Mother – Non-Hodgkin's lymphoma, 56; Maternal aunt – Leukaemia, 60; Maternal aunt – Breast, 65	Nil

Coefficient describes effect and magnitude of effect on gene expression of variants in eQTL (range -1.29 – 1.15 amongst higher confidence eQTL reported in cancer tissue

Table 6.31 - Single nucleotide variant affecting eQTL where variants reported to reduce *ZNF284* expression (heterozygous)

Transcript	Coordinate	Reference allele	Alternate allele	eQTL start	eQTL end	Coefficient	Distance to transcription start site (bp)	Phenotype with age at diagnosis	Family history	Clinically relevant coding variants detected
ENST00000270077	chr19:43772519	C	G	43772478	43772537	-0.99	-803790	Hemangiopericytoma, 51; Breast, 53	Other – Breast, 31 and oral cancer, 31.	Nil

Coefficient describes effect and magnitude of effect on gene expression of variants in eQTL (range -1.29 – 1.15 amongst higher confidence eQTL reported in cancer tissues)

6.7.4 - Putative proto-oncogenes, genes associated with telomere function, ultra-conserved regions or expression quantitative trait loci reported by GTEx project (see methods in 6.3, 6.4, 6.5.2.2 and 6.5.2.3)

Case-control comparisons were performed as described. No comparisons with q value <0.05 were noted at gene/region or variant level in any analysis including those incorporating counts of SNVs and indels, structural variants or the sum of both.

6.7.5 - Analysis for causative variants in a family with suspected recessive tumour predisposition (see methods in 6.6)

For the homozygous hypothesis, two variants passed filters (Table 6.32). The inframe deletion in *MSH3* affects a mismatch repair gene, a number of which are associated with Lynch syndrome and constitutional mismatch repair syndrome. Allele frequency of this variant is low in European populations (1000 Genomes 0.003, gnomAD 0.01) but is observed at a maximum of 0.34 in the gnomAD South Asian population and was not considered further on this basis.

ARVCF ENST00000263207 c.1616G>A (p.R539Q) does not occur in any gnomAD or 1000 Genomes population at a frequency above 0.01. Combined Annotation Dependent Depletion (CADD) score (phred scaled) for the variant is 31. 12 homozygotes are observed in gnomAD but this dataset contains TCGA data. 1 homozygote was observed in both the BRIDGE control series used for case control analyses and the 1958 birth cohort. A homozygote was also observed in the MPT series who had previously been diagnosed with bilateral phaeochromocytoma at the age of 59 years.

Table 6.32 - Variants passing filters according to a homozygous hypothesis

Chromosome	Position	Consequence	Gene	Transcript	Description
22	19965563	Missense	*ARVCF*	ENST00000263207	c.1616G>A (p.R539Q)
5	79950699	Inframe deletion	*MSH3*	ENST00000265081	c.154delGCAGCGGCTGCAGCGGCC (p.154del6)

One variant pair was identified by the filtering designed to identify compound heterozygote variants (Table 6.33). *COL6A2* ENST00000300527 c.679G>A p.(D227N) was identified in the father and c.988G>A (p.D330N) in the mother. Phred scaled CADD scores were 16.95 and 33 for each variant respectively.

Table 6.33 - Variants passing filters according to a compound heterozygous hypothesis

Chromosome	Position	Consequence	Gene	Transcript	Description
21	47532456	Missense	*COL6A2*	ENST00000300527	c.679G>A p.(D227N)
21	47536717	Missense	*COL6A2*	ENST00000300527	c.988G>A (p.D330N)

6.8 Discussion

Case control based analysis of a number of variant modalities and phenotypic subgroups revealed few loci where the evidence was indicative of a role in tumour predisposition.

Results that proposed causative loci with greatest confidence arose from truncating variant analysis, where variants in known CPGs contributed to results crossing the chosen significance thresholds. Consequently, there is considerable overlap between the results of this analysis and the WGS-based comprehensive CPG analysis described in a Chapter 4. For counts of individuals with truncating variants per gene, *CHEK2*, *PALB2*, *MAX* and *NF1* were over-represented in various phenotypic subgroups. Occurrence of individuals with two specific variants in *PALB2* and *CHEK2* was significantly higher in one subgroup each. The appearance of these results involving known CPGs indicates that the experimental design was able to propose regions in which constitutional variants cause susceptibility to neoplasia.

Genes producing top results in the analysis have characteristics (apart from being CPGs) leading to a greater probability of variants affecting them appearing in pre-assessed clinical genetics referral-based series. At the point of consultation for most participants, none of them were routinely tested. This is in contrast to genes such as *BRCA1* where molecular diagnosis would likely have been made in the clinic and study recruitment not undertaken. *CHEK2*, and *PALB2* are well established as being associated with breast cancer predisposition but uncertainties regarding penetrance or clinical utility of testing have previously inhibited frequent molecular investigation. *MAX* is a relatively recently described CPG and the individuals harbouring truncating variants affecting it would likely be detected by clinical services if presenting now. Neurofibromatosis type 1 is commonly seen in genetics clinics but has historically been a largely clinical diagnosis with molecular testing of *NF1* generally not performed.

Phenotypic subgroups producing top results generally contained tumours that were characteristically associated with the relevant gene. Although some subgroups appeared to suggest novel tumour types arising from variants in particular genes, further delineation of the phenotype of cases contributing to those results revealed they had also been diagnosed with typical tumours e.g. *PALB2* in individuals with ≥1 tumour from haematological myeloid, aerodigestive tract, anal and melanoma. The only result reaching the chosen significance threshold in the pan-cancer analysis of all cases arose from comparison of the count of individuals with *CHEK2* truncating variants. Variants in *CHEK2* have been associated with a wide variety of cancers[206,385] and the most robust of these associations is with breast cancer.[42,198,386] Consistent with this, most individuals with *CHEK2* truncating variants had previously been diagnosed with that tumour. Other associations might be suggested by non-breast tumours occurring in variant carriers (where an individual may or may not have had breast cancer).

Only RCC appeared with sufficient frequency to suggest an association here but was not significantly over-represented in truncating variant carriers vs non-carriers. Further research to investigate a possible relationship to the development of RCC include larger studies of variant carriers (including non-c.1229delC/c.1100delC) or tumour studies such as loss of heterozygosity analysis.

When counts of individuals with single nucleotide variants or indels were considered in combination with SV counts, *BMPR1A* variants were over-represented amongst all MPT cases and in various subgroups involving colorectal cancer although this result was due to only two individuals, one of whom did not have colorectal cancer and had an SV. This was a translocation with a breakpoint within the gene but between exons 1 and 2, which are both non-coding. Additionally, review in IGV suggested this SV may be an artefact due to multiple alignments of supporting sequencing reads. *BMPR1A* is associated with Juvenile Polyposis and colorectal cancer and the other individual with the SNV in this gene (nonsense) had previously been diagnosed with the latter. They have previously been described in chapters 4 and 5 and also harboured a truncating *PMS2* variant. Taking these factors into consideration, there is no evidence for a novel CPG locus from this result and little evidence for a novel phenotype caused by *BMPR1A* variants.

The other gene highlighted by incorporation of SVs into variant counts was *HABP2*, where one individual had a predicted translocation with a breakpoint between (coding) exons 1 and 2 and three other individuals had nonsense variants. These individuals contributed to highlighted results in two phenotypic subgroups incorporating a wide variety of tumour types that were assembled to emulate those seen in Peutz-Jeghers Syndrome and Li Fraumeni Syndrome. Indeed, apart from breast cancer shared between two of these participants the neoplastic manifestations in them were disparate. *HABP2* is a serine protease,[380] variants in which are associated with susceptibility to non-medullary thyroid cancer due to a report concerning a missense variant in single family that proposed *HABP2* as a tumour suppressor gene.[387] No thyroid cancers were reported in the individuals with *HABP2* variants in the currently presented MPT series. This gene is somatically mutated in a large proportion of some cancer types in cBioPortal, including colon cancer, but these percentages result from small sample sizes.[34] Expression data from the cancer genome atlas does not indicate under-expression of *HABP2* in TCGA provisional datasets relevant to the tumours observed in variant carrying individuals (mean z-score between -0.04 and 0.07 for invasive breast carcinoma (n=1100),[388] colorectal adenocarcinoma (n=382)[336] and phaeochromocytoma/paraganglioma (n=184)[389]). Given the marginal difference in the proportion of cases and controls with *HABP2* variants and lack of further information from investigation of participants tumours, it cannot be concluded that disruption of this gene was relevant in the causation of the neoplasms observed.

All other results highlighted by a q-value of <0.05 arose from analysis of variants in eQTL previously reported in cancer tissues but variant counts were small and causative effects appeared unlikely based on other lines of evidence. Four individuals were observed with a single eQTL variant (chr7:141437957 T>C) reported to reduce *TAS2R5* expression (in tumours including colorectal cancer), leading to significant results in phenotypic subgroups incorporating colorectal and/or aerodigestive tract cancer. Notably, two of the individuals with colorectal cancer carried a homozygous pathogenic variant in *NTHL1*, which is likely to have caused their tumours although a modifying effect of the eQTL variant is feasible. *TAS2R5* encodes a bitter taste receptor,[380,390] a function that is unlikely to be related to neoplastic processes and its product does not have any physical interactions with known CPGs in Gene Mania.[346] *TAS2R5* is not significantly under-expressed in colorectal adenocarcinoma (n=382, mean z-score 0.2)[34,391] or head and neck squamous cell carcinoma TCGA studies (n=521, mean z-score 0.03).[34,392] It is mutated or deleted in <1% of those cancer types in cBioPortal and the most frequent aberration is amplification in 21% of prostate cancers.[34]

Variants in an eQTL reported to upregulate *ENPP2* expression (in tumours including ovarian serous cystadenocarcinoma) were observed in 37.5% of individuals with both breast and ovarian cancer compared with 14% of controls. Mechanistically, *ENPP2* has a number of functions to suggest a role in tumourigenesis as it has been shown to promote angiogenesis and tumour cell motility. Its expression is upregulated in various carcinomas[380,390] but no over-expression is reported in the TCGA breast invasive carcinoma (n=1100, mean z-score 0.12)[34,388] or ovarian cystadenocarcinoma (n=307, mean z-score 0.09).[34,393] Amplification, however, is observed in 20% of ovarian and 10.5% of breast cancers (as well as 40.3% prostate cancers) in cBioPortal.[34] There are some indications, therefore, that further studies of variants at this eQTL in breast-ovarian cancer cases may be rewarding. These might include assessing their frequency in larger cohorts or analysing *ENPP2* expression in the tumours of individuals found to carry variants. However, caution should be drawn from the observation of multiple indel alleles at matching or nearby sites contributing to the results in this analysis, which may indicate sequencing or variant calling error.

Two individuals with variants in an eQTL reported to lead to reduced expression of *C2orf27A* were sufficient to produce results with q-values below the significance threshold for cases with both breast and kidney cancer. The effect on expression was not noted in breast or kidney cancer in the original publication, no under-expression of *C2orf27A* is noted in these tumours in the relevant TCGA studies[388] (Breast invasive carcinoma TCGA provisional, Kidney Renal Papillary Cell Carcinoma TCGA Provisional[394] and Kidney Renal Clear Cell Carcinoma TCGA Provisional[395]) and the gene is deleted, amplified or mutated in <1% breast or kidney cancers in cBioPortal.[34] There is little known about the function of *C2orf27A*.[380] Furthermore, inspection of BAM files from the two individuals

with variants in IGV showed a majority of reads at this region being of low mapping quality, bringing the variant call into doubt.

Combination of counts of individuals with SNVs/indels or SVs affecting eQTL and comparison in cases and controls produced one result where the q-value was below the significance threshold and where both SNVs/indels and SVs contributed to the result. This was contributed to by variants in an eQTL reported to reduce *ZNF284* expression in breast invasive carcinoma, colon adenocarcinoma/rectum adenocarcinoma, ovarian serous cystadenocarcinoma, liver hepatocellular carcinoma and lung adenocarcinoma. Two individuals amongst cases with at least one tumour from colorectal, breast, gastric or ovary sex cord-gonadal stromal (Peutz-Jeghers like) had a variant affecting this eQTL, although one of these was a predicted deletion where review in IGV offered little evidence to confirm or refute the variant. *ZNF284* encodes a zinc finger protein with nucleic acid binding properties.[380] Considering the tumours that occurred in the two variant carrying individuals, the gene is mutated in 2.28% of 439 colorectal adenocarcinomas and less than 1% and soft tissue sarcomas in cBioPortal. Higher aberration rates are seen in prostate cancers but this refers to amplification rather than mutation or deletion as would fit with the proposed mechanism here.[34] mRNA expression data from TCGA provisional studies does not indicate under expression in any of the cancer types observed in these individuals in terms of mean z-score.[336,388,396,397] Taken together with the small number of variant carrying cases without a shared phenotype, these lines of evidence do not suggest that further investigation of this locus would be rewarding in this context.

Analysis of a family with possible recessive inheritance of predisposition to osteomas resulted in a homozygous variant in *ARVCF* for further consideration as well as a pair of *COL6A2* variants under a compound heterozygous hypothesis. *ARVCF* is located in the 22q11 deletion region associated with a developmental syndrome (heterozygous deletions) primarily causing congenital heart disease, cleft palate, learning difficulties and immunodeficiency rather than neoplasia. There is little suggestion of phenotypic overlap with that syndrome in the studied family but *ARVCF* is a member of the catenin family and is involved in the formation of adherens junction complexes. A recognised CPG that shares this function is *CDH1*, variants in which are associated with hereditary diffuse gastric cancer and lobular breast carcinoma.[96,398] ARVCF interacts with the *CDH1* gene product e-cadherin and the ARVCF domain that this variant occurs within has been reported to be necessary for binding between the two proteins.[399] A single submission of this variant in ClinVar reports the variant as benign but no phenotype for which this assertion is made is given. cBioPortal was interrogated for variants in *ARVCF*, which is mutated in around 17% of central nervous system tumours but this figure is contributed to by a single case only. A limitation of this query is the fact that, as for a number of rarer and/or benign tumours, there is no osteoma study available via that platform. Validation cohorts with similar phenotype are often crucial to CPG discovery. The recent elucidation of *POLE* and *POLD1* as

polyposis/colorectal cancer predisposition genes initially identified variants in them in a single family but further occurrences in a cohort of individuals with colorectal cancer (and absence in controls) indicated a causative effect.[35] Despite the functional evidence suggesting a possible role for this variant in neoplasia, there is little else to suggest a causative role in the studied family and there was no apparent phenotypic overlap with the other homozygote in the MPT series. Similarly, there was insufficient evidence that the compound heterozygous variants in *COL6A2* had a role in the development of osteomas in these individuals. *COL6A2* is a type VI collagen gene, pathogenic variants in which are associated with myopathies[400] but not with neoplastic phenotypes. Entries in ClinVar exist for both variants but only with pathogenicity assertions relating to myopathy and no entry reporting either variant as pathogenic. SNVs/indels are not noted in any cancer type in cBioPortal at a frequency of 10% or more and no physical interactions between *COL6A2* and known CPGs are highlighted by the GeneMania platform.[346]

The paucity of novel loci highlighted as potentially causative by these analyses is likely due to a number of factors. Aside from truncating variants in known CPGs, the prior probability of causative variants in the analysed regions can be assumed to be low as to date, relatively few CPGs have been described in which variants lead to levels of tumour risk amenable to genetic counselling and risk mitigation. In the case of non-coding variants, prioritisation of variants and regions likely to be relevant to disease states is less developed than for coding regions. In this project, ultra-conserved regions, enhancers, promoters and eQTL were used. These resources are likely to be expanded and refined with time and other strategies are likely to improve estimation of non-coding variant pathogenicity. For example, the FUN-LDA tool utilises epigenetic information from large epigenetic data sources to assess likelihood of the significance of a genomic region to gene expression in a tissue specific manner and prioritise variants accordingly.[401] Functional assays are also likely to be an important tool and can be designed as high throughput techniques to maximise information obtained regarding the impact of induced variants. Strategies include the generation of multiple plasmids with different variations in putative regulatory regions and the observation of their effect on transcription via transfection and reporter assay. CRISPR-Cas9 based systems have also been used to generate multiple cell lines with distinct regulatory region variants with subsequent observation of the chosen phenotypic effect.[217]

To minimise false positives amongst the results, stringent filtering for genotype quality, sequencing depth and variant allele fraction was used. These measures were deemed necessary to avoid variant calls due to sequencing artefact but may have excluded some genuine variants. Given that a small number of rare variants can produce a low q-value in analyses such as those undertaken here, some overrepresented variants or regions may not have been highlighted.

Power of case-control based analyses is enhanced by large numbers of cases with a specific commonality between them. The identification of *PALB2* as a CPG was based on the observation of ten truncating variants in 923 breast cancer cases. These were familial, enhancing the probability of constitutional predisposing factors being present.[39] *NTHL1* was reported as predisposing to colorectal polyps and cancer through exome sequencing of samples from a lower number (n=51) of individuals but all had the relatively specific phenotype of multiple colorectal adenomas (48/51 had >10 recorded).[36] *MAX* was discovered as a CPG using exome sequencing on samples from only three individuals with phaeochromocytoma but these cases were familial and phaeochromocytoma is a highly heritable neoplasm.[286] The largest phenotypic group defined in these analyses was that comprising all MPT cases fulfilling inclusion criteria (n=424) though this subset was highly heterogeneous in terms of diagnosed tumours. Phenotypic subgroups were defined to decrease heterogeneity but this led to significant decreases in the number of individuals included in each group. For example, the largest group defined by a single tumour type was breast cancer, which included 215 participants. The largest group defined by a specific tumour combination was breast-colorectal but only 42 individuals were included. Although some results highlighted likely causal relationships with low numbers of participants (e.g. *NF1* truncations in 15 GIST cases), others may have remained undetected.

An alternative strategy to identify candidate causal variants that does not require a large number of probands is segregation analysis within families according to a hypothesised mode of inheritance. Recessive inheritance was proposed in a family where two siblings, born to unaffected parents, developed osteomas in childhood. Although it cannot be concluded that the filtered variants are causative, the process highlights the ability of the technique to efficiently narrow down candidates. The MPT series is largely composed of probands and although participants were contacted with the aim of recruiting family members, the cohort contained no other families where data from both parents was available and more than one individual was affected. Segregation based analysis could therefore not be performed in multiple families, which may have yielded positive results.

Discrepancies between estimated heritability of cancer types and the proportion of cases explained by known constitutional genetic factors[402,403] suggest that continued investigation may yield novel CPGs, although this can be stated with lower confidence for rarer tumour types without a robust heritability estimate. Missing heritability, however, does not necessarily imply a significant role for high penetrance variants in single genes and a proportion can be accounted for by more common, lower penetrance variants identified through genome wide association studies.[402,404] Under a polygenic risk model, co-occurrence of such variants in an individual may confer additional risk and scores to assess risk based on the burden of selected risk variants have been previously applied to investigate their clinical utility.[43,44,405,406] Missing heritability may also be accounted for by modalities of variation that

are not typically considered in studies to identify predisposing factors, most obviously non-coding or epigenetic variation.

Chapter 7 – Reflections and future perspectives

This research applied massively parallel sequencing techniques, in particular whole genome sequencing (WGS) to a series of individuals with multiple primary tumours (MPT). MPT was taken as an observation indicating an increased probability of a cancer predisposition syndrome due to a constitutional deleterious variant in a cancer predisposition gene (CPG).

Investigations were undertaken to elucidate causative variants affecting known CPGs that would be of immediate clinical relevance. A key finding of these analyses was that the use of MPT (as defined by the study eligibility criteria) *per se* as an indication for application of agnostic genetic testing would yield a substantial number of variants associated with clinical utility. Occasionally multiple such variants would be revealed in the same individual. The detection rate is enhanced by the use of WGS due to its ability to detect structural variants and interrogate any region of interest but these advantages are limited at present. They are likely to become more prominent as the cost of WGS decreases and greater characterisation of clinically relevant non-coding regions takes place.

7.1 - Variant assessment

Defining phenotypic effects caused by non-coding variation is a developing field but interpretation of coding variants in the context of human disease also remains challenging. In the assessment of variants for clinical relevance in this project, multiple exclusions were made on the basis of insufficient evidence leading to variant of uncertain significance (VUS) classification. This is a prominent issue in clinical and research settings. A large amount of work has previously been undertaken to improve the situation and define the risks associated with individual CPG variants.

A recent advancement, used extensively in this project, has been to build on previous efforts and formulate a consensus as to what lines of evidence should be used to assign pathogenic or benign status, including how each of them should be weighted. The American College of Medical Genetics (ACMG) guidelines[192] have been widely adopted but this body recognises that the criteria leave room for ambiguity as to whether a threshold should be crossed for a given line of evidence. For example, what functional assays qualify as "well established" and what result of that assay can be taken as evidence of a damaging effect? A study involving four diagnostic laboratories in the United States observed pathogenicity assessments (not using ACMG criteria) of any variant that had been submitted to ClinVar by two or more of them. 242 discordant variants were reassessed by the respective laboratories using the ACMG criteria but a 12.8% discordance rate remained.[407] A response to inconsistencies such as these has been to form working groups that apply guidelines in a manner specific to the disease or gene in question. A published example of this approach is the refinement of ACMG criteria application in the context of *MYH7*-associated inherited cardiomyopathies.[408] Here, nine criteria were deemed not applicable and clarifications were made regarding aspects such as

degree of segregation considered sufficient to support pathogenicity. For cancer predisposition syndromes, the UK Cancer Genetics Group is undertaking a similar process.

Work to improve the ACMG guidelines may also focus on individual criteria rather than diseases or genes. To this end, a working group of clinical and laboratory geneticists was assembled to discuss application of the criterion fulfilled if the assessed variant has a predicted loss of function consequence (PVS1).[409] Prior to this, the ACMG had issued a recommendation that the weighting assigned to a fulfilled criterion (very strong, strong, moderate or supporting) could be shifted on the basis of further evidence (e.g. a supporting line of evidence could become strong).[410] Recommendations from the working group included consideration of whether a nonsense variant affects a biologically relevant transcript, the proportion of a protein lost as a result of a variant and, in the case of splice variants, the presence of nearby consensus splice sequences that may re-establish in frame splicing. In this project, the former two aspects were taken into consideration in predicted loss of function variant assessment but guidance such as this will promote consistency in future assertions of pathogenicity.

Aside from consensus, work to provide further evidence as to the phenotypic effects of constitutional sequence variants is ongoing. The ClinVar database continues to expand and now exists in partnership with the ClinGen programme[411] to enhance expert curation in terms of whether genes are associated with a given disease, whether variants are pathogenic and what clinical action can be taken as a result of their detection. The array of in-silico tools to predict variant consequences continues to grow and can be improved by expanded variant databases on which to base algorithms. A recent example is ClinPred,[412] which formulates a score based on existing in-silico tools (e.g. Combined Annotation Dependent Depletion (CADD)) as well as allele frequency information from the gnomAD dataset. ClinVar variants were used as a training dataset, which was considered to be superior to other curated variant databases due to its size and pathogenicity assertions based on ACMG criteria.

A valuable source of information regarding the phenotypic consequence of a variant is the results of functional assays designed to observe its effect in a model system. Execution of these experiments is laborious and evidence relating to individual variants is frequently unavailable but higher throughput techniques are being utilised that have the potential to dramatically expand the range of variants for which functional studies have been undertaken. A notable recent report analysed the effect of 1,056 *BRCA1* missense variants on repair of double stranded breaks (DSBs) by homologous recombination.[413] A cell line was utilised where effective DSB repair is observed through expression of a *GFP* gene from a genomic insertion designed with a target site for a transfected DSB inducing enzyme. A second inactive (due to an absent promoter) copy of *GFP* was also included in the insert and used as a template for repair if that process was functional. Multiplexed reporter assays and

mutagenesis to generate plasmids for them allowed the high number of variants to be generated and analysed.

7.2 - Atypical phenotypes

A further key finding of this project was the high rate of tumour types in pathogenic/likely pathogenic CPG variant carriers that were not characteristically associated with disrupted function of the relevant CPG. In the presented analysis, around 40% of studied probands had been diagnosed with at least one atypical tumour. The preferential consideration of MPT cases is likely to elevate this figure but other (non-MPT based) reports also report high rates of discordant neoplasms.[200,203] These observations are becoming more frequent as genetic testing is more broadly applied and a significant challenge is to distinguish incidental tumours from those which have been contributed to by the identified constitutional CPG variant. Functional assays are less likely to be helpful in answering this question as results from cell lines cannot be readily extrapolated to *in vivo* tumour subtypes. Variant databases such as ClinVar give valuable information as to the pathogenicity of a variant in the context reported by the submitter but do not provide numerical risks of tumours as a result of the variant.

Success in defining the risks associated with variants in CPGs has been achieved by collating variant carriers in a manner that seeks to minimise ascertainment biases. Examples include a prospective study of carriers of pathogenic mismatch repair gene variants[69] and an analysis of succinate dehydrogenase subunit gene variant carriers that considers the rate of phaeochromocytoma/paraganglioma in relatives testing positive through predictive testing.[414] These strategies are generally focused on recording tumours already known to be associated with variants in the studied syndrome or gene but are also well placed to highlight novel associations. A difficulty is collating sufficient numbers of individuals with rare CPG variants in a given gene but an initiative to address this issue is proposed as part of the "Cancer Moonshot" initiative by the National Cancer Institute. A "Pre Cancer Genome Atlas" is planned that aims to assemble CPG variant carriers identified through genetic testing and create an information sharing platform from their data.[415]

7.3 - Identifying novel loci relevant to tumour predisposition

A number of interrogations of WGS data were made as part of this research that aimed to identify novel loci associated with tumour predisposition. These were predominantly based on defining putative regions of relevance (e.g. genes recurrently somatically mutated in cancer, ultra-conserved regions, gene enhancers) and comparing the frequency of variants within them in various phenotypically defined case groups vs controls. Truncating variants in some genes (*NF1*, *PALB2*, *MAX*, *CHEK2*) were found to be over-represented in some case groups vs controls, illustrating the potential efficacy of this approach. However, these results did not represent novel CPG loci or robust

gene-tumour phenotype associations and other interrogations did not produce convincing evidence of causative variants.

Studies utilising massively parallel sequencing data to identify disease associated variants frequently generate large numbers of potentially causative variants that require prioritisation based on one or more lines of evidence to reduce the number of candidates. In each of the analyses for novel loci, prioritisation was undertaken in this manner but the number of candidate regions remained large in many instances, resulting in a large number of tests informing calculations to adjust p-values in light of multiple hypotheses. Some pertinent results might not have crossed chosen statistical significance thresholds for this reason and further information to narrow candidate regions may have avoided this.

A further limitation of attempts to elucidate novel loci in this project was the phenotypic heterogeneity of studied individuals. The over-representation of variants in a particular region that is associated with a phenotype is more likely to be detected where the phenotype that defines the case group is more specific. An increase in heterogeneity will dilute any cases with a shared genetic cause and may lead to pertinent results not being highlighted through hypothesis testing. The analyses undertaken here took steps to increase phenotypic specificity through subgrouping individuals by tumour type but this led to a large reduction in the number of cases in most subgroups, itself a cause of failure to detect regions in which variants were over-represented.

7.4 - Tumour sequencing

An undertaking that has the potential to address many of the difficulties highlighted above is the expansion of tumour (as well as concurrent germline) sequencing in diagnostic settings and resultant use of generated data/information in research contexts. This practice promises to enhance the identification and collation of CPG variant carriers as well as provide molecular data that could assist with variant interpretation, defining phenotypic subgroups for research and prioritising putative candidate regions containing tumour predisposing variants.

Genotyping of tumour samples is frequently undertaken as part of cancer management but the testing is usually narrow in scope and designed to detect variants in specific genes that will inform prognosis and/or treatment of that cancer type. Examples include analysis for *HER2* amplification in breast cancer that would prompt Trastuzumab therapy and *EGFR* mutations in non-small cell lung cancer to guide the use of Afatinib. Next generation sequencing assays provide the opportunity to perform most of the tests in current clinical use with a single assay that would also generate data to identify other useful markers or apply existing markers in other tumours. The widespread use of this strategy has been advocated in the Chief Medical Officer's 2016 "Generation Genome" report[416] and

establishment of workflows for routine WGS of tumours is a key aim of the 100,000 Genome Project, an initiative that also analyses constitutional DNA from blood samples.

Enhanced sequencing of tumours as part of routine care pathways has great potential to increase identification of individuals harbouring constitutional pathogenic CPG variants. One adopted strategy might be to use a gene panel assay of cancer driver genes, which overlap extensively with CPGs. Detection of a variant that could be relevant to tumour predisposition could prompt analysis of a blood sample to assess somatic vs germline status and indicate whether referral to genetics services was appropriate. A more comprehensive approach would be to routinely perform extensive sequencing on both tumour and non-tumour (e.g. adjacent normal or blood) tissue, allowing a number of inferences to be made. Presence of a CPG variant in a tumour but not the germline would indicate a somatic variant but the possibility of mosaicism would need to be considered if other tumours from the same individual contained the same variant or if it was detectable at a low variant allele fraction in blood. Detection of a CPG variant in tumour and blood may indicate a possible cancer predisposition syndrome and further assessment by genetics services would be indicated. Further useful information as to the variant's role in the development of the tumour might be obtained if loss of the wild type allele was demonstrated in the tumour, an observation that would require reasonable sequencing depth to make with confidence. A further scenario is the identification of a CPG variant in a blood sample but not in the tumour. This might imply that it was not significant to tumourigenesis in the sample in hand but does not necessarily provide reassurance that the individual or their family are not at risk of other tumours. Furthermore, constitutional variants can be lost in tumours. In a study of 198 advanced cancer cases with pathogenic assessed CPG variants identified through tumour-normal sequencing (of 341 genes), 13 had a monoallelic CPG variant lost in the tumour sample.[200] Loss of variant alleles may occur through genomic instability and not be relevant to tumour progression but the possibility also exists of a variant that is important for tumour initiation (i.e. acting as a pathogenic CPG variant) but incompatible with survival of the later neoplastic clone.

Another possible future mechanism whereby carriers of constitutional pathogenic CPG variants might be detected is through population screening. It would be feasible for this to take the form of directly sequencing germline DNA samples to detect rare deleterious CPG variants. This idea has been discussed for *BRCA1* and *BRCA2* screening, particularly in Ashkenazi Jewish populations where prevalence is higher. The approach would certainly reduce ascertainment biases influencing risk estimates surrounding CPGs but may come with unacceptable costs in terms of economics and negative impact of variant detection such as psychological distress or prophylactic surgery in lower risk carriers.[417–419] An intriguing alternative form of population screening that might identify unaffected individuals with cancer predisposition syndromes is analysis of circulating tumour DNA (ctDNA), in effect performing tumour sequencing without prior knowledge of a neoplasm to biopsy.

Many suggested applications of ctDNA based techniques relate to monitoring of drug response or recurrence but a recent study reported good sensitivity and specificity in detecting a range of non-metastatic (but clinically detectable) common cancers through a test utilising ctDNA in combination with protein biomarkers.[420] This "CancerSEEK" test is proposed as a possible basis for future population screening for common cancers rather than cancer predisposition syndromes but there is frequently overlap between genes containing known somatic driver mutations and CPGs. CancerSEEK generates sequence data for 16 genes and five of these are known CPGs (*APC*, *EGFR*, *PTEN*, *TP53*, *HRAS*). In the CancerSEEK study, detection of a potential cancer driver mutation in plasma DNA prompted interrogation for the same variant in lymphocyte DNA to exclude it if present in the germline. However, presence of CPG variants identified in this way that were also present in lymphocytes might prompt further assessment of the individual for a predisposition syndrome. Notably, the test only utilised 61 amplicons to assess common driver mutations rather than sequence large areas of genes but future assays might broaden the sequence information generated.

Apart from identifying CPG variant carriers, expansion of tumour sequencing would also assist with clinical decision making and research projects in other ways depending on the assay performed.

Although not applicable to all tumour predisposition syndromes (e.g. those due to gain of function variants in proto-oncogenes), demonstration of loss of the wild type allele in a tumour sample where a constitutional CPG variant is present provides evidence of a role for that variant in tumourigenesis, assuming a tumour suppressor gene two-hit model. Such loss of heterozygosity (LOH) may be due to deletion of the wild type allele or be copy number neutral due to somatic uniparental disomy. It may also occur through mutation of the wild type allele or due to an epimutation, the latter of which is not detectable without specialised sequencing techniques.

In clinical practice, the use of LOH analysis in variant assessment is well established but often performed on a *post hoc* basis for specific variants and tumour tissue is frequently unavailable. More use of routine tumour sequencing in diagnostic laboratories would provide greater opportunity to interrogate regions corresponding to putative pathogenic CPG variants to observe the relative allelic ratios in a tumour vs blood sample. Detection of LOH may not require WGS or exome sequencing of tumours and necessary data could be obtained through potentially cheaper assays designed for other purposes. The recently reported Karyogene assay was developed for myeloid malignancies (i.e. not for concurrent solid tumour-blood sequencing) but illustrates how a diagnostic assay may reveal LOH for a detected CPG variant. Here, high depth sequencing was based on capture by a series of oligonucleotide baits targeting exons of genes of interest, breakpoints of known translocations and also single nucleotide polymorphisms (SNPs) every 300kb.[421] The latter is used to identify regions of

homozygosity to detect copy number variants causing myeloid malignancies but could be used for other applications.

In research studies attempting to identify novel tumour suppressor CPGs in constitutional DNA, regions of homozygosity identified in tumour through WGS, exome sequencing or (lower cost) SNP based approaches could be used to narrow candidate regions. The reduction in putative causative variants from this approach may allow analytical time and resources to be better focused.

Genetic analysis of tumours has most frequently focused on specific regions (such as genes) where variation can be interpreted as having a biological effect. This kind of analysis can be performed with sequence data covering a relatively small area of the tumour genome but more expansive techniques such as WGS can observe accumulated variants across all the genome. Consequently, a picture of the mutational processes that have taken place can be obtained with commonalities and differences between neoplasms analysed. These mutational signatures have been conceptualised and defined in recent years and can reflect the known environmental exposures relevant to specific cancer types (e.g. higher rate of C>T mutations in melanoma due to nucleotide excision repair of ultraviolet induced pyrimidine dimers).[281] They can also demonstrate underlying tumourigenic genetic abnormalities, which may produce a contrasting signature to that usually seen in a given tumour type. For example, mismatch repair deficiency can be identified and a characteristic pattern of indels is observed in breast cancers from individuals with pathogenic *BRCA1* or *BRCA2* variants, which is taken to indicate deficient double stranded break repair by non-homologous end joining.[422,423]

Mutational signatures in tumours undergoing WGS by diagnostic services could be exploited in a number of ways. In the clinic, the presence or absence of a signature associated with deleterious variants in a particular gene could be used to infer the pathogenicity status of a constitutional variant in that gene following identification in a blood sample. In research settings, signatures could be used to define participant subgroups and enhance phenotypic specificity. Strategies might include excluding individuals whose tumours show a typical or environmental exposure related signature. Additionally, research participants could be grouped according to a common mutational signature in their tumours that may be unexplained and/or be present in neoplasms from multiple anatomical sites or tissues.

The analyses of tumour genomes described above are reliant on good quality DNA in sufficient quantity that has been extracted from tumour tissue. Historically, tissue obtained through biopsy or surgical resection has been fixated using formalin and embedded in paraffin. This has served pathologists well as structures are preserved for microscopy and samples can be easily stored at room temperature. However, formaldehyde interacts with DNA through a number of chemical reactions that

can lead to sequencing artefacts via disruption of DNA polymerases used in polymerase chain reactions.[424] Cross linking of nucleic acids and proteins induced by formaldehyde induces fragmentation[425] that can compromise DNA library preparation. Nucleic acids extracted from formalin fixed paraffin embedded (FFPE) tissue are frequently used for genetic analysis but the probability of obtaining a high-quality result declines as the scope of the test increases. WGS requires higher yields of DNA and is unlikely to be successful using FFPE tumour samples.

These issues can be overcome by the use of fresh frozen samples as this process does not induce the reactions and cross linkages associated with formalin. Practical difficulties with frozen tissue include the necessity to freeze the sample quickly after removal from the patient and storage in freezers rather than at room temperature. Nevertheless, a transition towards this procedure in surgical departments and histology laboratories is necessary if the full potential of cancer genomic medicine is to be realised. The 100,000 Genomes Project cancer arm has taken the decision to only accept fresh frozen tissue for sequencing apart from in exceptional circumstances.[426] As a major aim of project is to establish optimal workflows for genomic medicine in healthcare settings, it is hoped that this initiative will pave the way for extensive sequencing of cancer tissues to be performed routinely.

Cancer predisposition syndromes, although not common, represent a good target for high impact preventative strategies given the level of risk they frequently confer and the potentially severe consequences of neoplastic disease. Work to characterise them, identify them in patients and mitigate risk have led to significant benefits for affected individuals due to extensive work over a long period of time. This work has hitherto been restricted by limitations in capability to sequence patient samples but advancements in this area have begun to lift them. The combination of accumulated knowledge and application of genomic technologies offers great opportunities in the continuation of efforts to improve risk estimation and preventative strategies for those at increased risk of cancer.

References

1. Whitworth J, Maher ER. Cancer Genetics and Genomics. In: Kumar D, Antonarakis S, eds. *Medical and Health Genomics*. 1st ed. Elsevier Inc.; 2016:261-285.
2. Wunderlich V. JMM — Past and Present. *J Mol Med*. 2002;80(9):545-548.
3. Boveri T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J Cell Sci*. 2008;121 Suppl(Supplement_1):1-84.
4. Nowell PC. Review series personal perspective Discovery of the Philadelphia chromosome : a personal perspective. 2007;117(8):2033-2035.
5. Bishop J. Retroviruses and cellular oncogenes. *Annu Rev Biochem*. 52(1983):301-354.
6. Nordling CO. A new theory on cancer-inducing mechanism. *Br J Cancer*. 1953;7(1):68-72.
7. Martincorena I, Raine KM, Gerstung M, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2017;171(5):1029-1041.e21.
8. Knudson AG. Mutation and Cancer : Statistical Study of Retinoblastoma. 1971;68(4):820-823.
9. Yunis JJ, Ramsay N. *Retinoblastoma and Subband Deletion of Chromosome 13.* American journal of diseases of children (1911) 132, 161-163 (1978).
10. Balaban G, Gilbert F, Nichols W, Meadows AT, Shields J. Abnormalities of chromosome #13 in retinoblastomas from individuals with normal constitutional karyotypes. *Cancer Genet Cytogenet*. 1982;6(3):213-221.
11. Fung YT, Murphree AL, Tang A, Qian JIN, Hinrichs SH, Benedict WF. Structural Evidence for the Authenticity of the Human Retinoblastoma Gene. *Science (80-).* 1986;236(4809):1657-1661.
12. Mulligan LM, Kwok JB, Healey CS, et al. Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A. *Nature*. 1993;363(6428):458-460.
13. Donis-Keller H, Dou S, Chi D, et al. Mutations in the RET proto-oncogene are associated with MEN 2A and FMTC. *Hum Mol Genet*. 1993;2(7):851-856.
14. Schmidt L, Duh FM, Chen F, et al. Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas. *Nat Genet*. 1997;16(1):68-73.
15. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. 1953;171(4356):737-738.
16. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016;107(1):1-8.
17. Smith HO, Wilcox KW. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J Mol Biol*. 1970;51(2):379-391.
18. Lehman IR, Bessman MJ, Simms ES, Kornberg A. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli. *J Biol Chem*. 1958;233(1):163-170.
19. Bessman MJ, Lehman IR, Simms ES, Kornberg A. Enzymatic synthesis of deoxyribonucleic

acid. II. General properties of the reaction. *J Biol Chem*. 1958;233(1):171-177.

20. Jackson DA, Symons RH, Berg P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proc Natl Acad Sci U S A*. 1972;69(10):2904-2909.
21. Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol*. 1987;155:335-350.
22. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*. 1975;94(3):441-448.
23. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A*. 1977;74(2):560-564.
24. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology*. 1992;24:104-108.
25. Venter JC, Adams MD, Myers EW, et al. The Sequence of the Human Genome. *Science (80-)*. 2001;291(5507):1304-1351.
26. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376-380.
27. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-59.
28. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*. 2003;299(5607):682-686.
29. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet*. 2010;19(R2):R227-40.
30. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*. 2009;4(4):265-270.
31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
32. Raczy C, Petrovski R, Saunders CT, et al. Isaac: Ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*. 2013;29(16):2041-2043.
33. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
34. cBioPortal for Cancer Genomics. http://www.cbioportal.org/. Accessed April 8, 2016.
35. Palles C, Cazier J-B, Howarth KM, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet*. 2013;45(2):136-144.
36. Weren RDA, Ligtenberg MJL, Kets CM, et al. A germline homozygous mutation in the base-

excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nat Genet*. 2015;47(6):668-671.

37. Meijers-Heijboer H, Van den Ouweland A, Klijn J, et al. Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations: The CHEK2-breast cancer consortium. *Nat Genet*. 2002;31(1):55-59.
38. Seal S, Thompson D, Renwick A, et al. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet*. 2006;38(11):1239-1241.
39. Rahman N, Seal S, Thompson D, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet*. 2007;39(2):165-167.
40. Antoniou AC, Casadei S, Heikkinen T, et al. Breast-cancer risk in families with mutations in PALB2. *N Engl J Med*. 2014;371(6):497-506.
41. Easton DF, Lesueur F, Decker B, et al. No evidence that protein truncating variants in BRIP1 are associated with breast cancer risk: Implications for gene panel testing. *J Med Genet*. 2016;53(5):298-309.
42. Southey MC, Goldgar DE, Winqvist R, et al. PALB2, CHEK2 and ATM rare variants and cancer risk: Data from COGS. *J Med Genet*. 2016;53(12):800-811.
43. Mavaddat N, Pharoah PDP, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst*. 2015;107(5):djv036-djv036.
44. Evans DG, Brentnall A, Byers H, et al. The impact of a panel of 18 SNPs on breast cancer risk in women attending a UK familial screening clinic: A case-control study. *J Med Genet*. 2017;54(2):104-110.
45. Rahman N. Realizing the promise of cancer predisposition genes. *Nature*. 2014;505(7483):302-308.
46. Lynch HT, Fusaro RM, Lynch J. Hereditary cancer in adults. *Cancer Detect Prev*. 1995;19(3):219-233.
47. Lu C, Xie M, Wendl MC, et al. Patterns and functional implications of rare germline variants across 12 cancer types. *Nat Commun*. 2015;6:10086.
48. O'Brien JM. Environmental and heritable factors in the causation of cancer: analyses of cohorts of twins from Sweden, Denmark, and Finland, by P. Lichtenstein, N.V. Holm, P.K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki. N. *Surv Ophthalmol*. 45(2):167-168.
49. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA*. 2016;315(1):68-76.
50. Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer*. 2002;99(2):260-266.

51. Al-Tassan N, Chmiel NH, Maynard J, et al. Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet*. 2002;30(2):227-232.
52. Astuti D, Latif F, Dallol A, et al. Gene Mutations in the Succinate Dehydrogenase Subunit SDHB Cause Susceptibility to Familial Pheochromocytoma and to Familial Paraganglioma. 2001;3(Mim 605373):49-54.
53. Alston CL, Davison JE, Meloni F, et al. Recessive germline SDHA and SDHB mutations causing leukodystrophy and isolated mitochondrial complex II deficiency. *J Med Genet*. 2012;49(9):569-577.
54. Gatti R. Ataxia-Telangiectasia. In: Pagon RA, Adam MP, Bird TD et al., ed. *GeneReviews™ [Internet]*. University of Washington, Seattle; 2010.
55. Swift M, Morrell D, Massey RB, Chase CL. Incidence of cancer in 161 families affected by Ataxia–Telangiectasia. *N Engl J Med*. 1991;325(26):1831-1836.
56. Schneider K, Zelley K, Nichols K, Garber J. *Li-Fraumeni Syndrome*. (Pagon RA, Adam MP, Bird TD et al., ed.). Seattle: University of Washington, Seattle; 2013.
57. Barshir R, Hekselman I, Shemesh N, Sharon M, Novack L, Yeger-Lotem E. Role of duplicate genes in determining the tissue-selectivity of hereditary diseases. *PLOS Genet*. 2018;14(5):e1007327.
58. Harbour JW, Onken MD, Roberson EDO, et al. Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science*. 2010;330(6009):1410-1413.
59. Horsman DE, White V a. Cytogenetic analysis of uveal melanoma: Consistent occurrence of monosomy 3 and trisomy 8q. *Cancer*. 1993;71(3):811-819.
60. Abdel-Rahman MH, Pilarski R, Cebulla CM, et al. Germline BAP1 mutation predisposes to uveal melanoma, lung adenocarcinoma, meningioma, and other cancers. *J Med Genet*. 2011;48(12).
61. Popova T, Hebert L, Jacquemin V, et al. Germline BAP1 mutations predispose to renal cell carcinomas. *Am J Hum Genet*. 2013;92(6):974-980.
62. Vanharanta S, Buchta M, McWhinney SR, et al. Early-Onset Renal Cell Carcinoma as a Novel Extraparaganglial Component of SDHB-Associated Heritable Paraganglioma. *Am J Hum Genet*. 2004;74(1):153-159.
63. Ricketts C, Woodward ER, Killick P, et al. Germline SDHB mutations and familial renal cell carcinoma. *J Natl Cancer Inst*. 2008;100(17):1260-1262.
64. Tan M-H, Mester JL, Ngeow J, Rybicki L a, Orloff MS, Eng C. Lifetime cancer risks in individuals with germline PTEN mutations. *Clin Cancer Res*. 2012;18(2):400-407.
65. Vasen HF, Mecklin JP, Khan PM, Lynch HT. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Dis Colon Rectum*. 1991;34(5):424-425.
66. Umar A, Boland CR, Terdiman JP, et al. Revised Bethesda Guidelines for hereditary

nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst*. 2004;96:261-268.

67. Aarnio M, Sankila R, Pukkala E, et al. Cancer risk in mutation carriers of DNA-mismatch-repair genes. *Int J Cancer*. 1999;81(2):214-218.
68. Stoffel E, Mukherjee B, Raymond VM, et al. Calculation of risk of colorectal and endometrial cancer among patients with Lynch syndrome. *Gastroenterology*. 2009;137(5):1621-1627.
69. Møller P, Seppälä T, Bernstein I, et al. Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. *Gut*. 2017;66(3):464-472.
70. Antoniou a, Pharoah PDP, Narod S, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet*. 2003;72(5):1117-1130.
71. Kuchenbaecker KB, Hopper JL, Barnes DR, et al. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA*. 2017;317(23):2402-2416.
72. Giusti F, Marini F, Brandi ML. Multiple Endocrine Neoplasia Type 2. In: Pagon RA, Adam MP, Bird TD et al., ed. *GeneReviews™ [Internet]*. University of Washington, Seattle; 2013.
73. Kloos RT, Eng C, Evans DB, et al. Medullary thyroid cancer: management guidelines of the American Thyroid Association. *Thyroid*. 2009;19(6):565-612.
74. Seri M, Celli I, Betsos N, Claudiani F, Camera G, Romeo G. A Cys634Gly substitution of the RET proto-oncogene in a family with recurrence of multiple endocrine neoplasia type 2A and cutaneous lichen amyloidosis. *Clin Genet*. 1997;51(2):86-90.
75. Hofstra RM, Landsvater RM, Ceccherini I, et al. A mutation in the RET proto-oncogene associated with multiple endocrine neoplasia type 2B and sporadic medullary thyroid carcinoma. *Nature*. 1994;367(6461):375-376.
76. NHGRI: Breast Cancer Information Core. https://research.nhgri.nih.gov/projects/bic/. Accessed July 25, 2014.
77. Boyd L, Parkin DM. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *J Epidemiol Community Heal*. 2011;65:A143-A143.
78. Butow PN, Lobb EA, Meiser B, Barratt A, Tucker KM. Psychological outcomes and risk perception after genetic testing and counselling in breast cancer: a systematic review. *Med J Aust*. 2003;178(2):77-81.
79. Metcalfe KA, Liede A, Hoodfar E, Scott A, Foulkes WD, Narod SA. An evaluation of needs of female BRCA1 and BRCA2 carriers undergoing genetic counselling. *J Med Genet*. 2000;37(11):866-874.
80. Hamann HA, Smith TW, Smith KR, et al. Interpersonal responses among sibling dyads tested for BRCA1/BRCA2 gene mutations. *Health Psychol*. 2008;27(1):100-109.
81. Barrow P, Green K, Clancy T, Lalloo F, Hill J, Evans DG. Improving the uptake of predictive

testing and colorectal screening in Lynch syndrome: a regional primary care survey. *Clin Genet*. 2015;87(6):517-524.

82. Meijers-Heijboer EJ, Verhoog LC, Brekelmans CT, et al. Presymptomatic DNA testing and prophylactic surgery in families with a BRCA1 or BRCA2 mutation. *Lancet*. 2000;355(9220):2015-2020.
83. Esplen MJ, Madlensky L, Butler K, et al. Motivations and psychosocial impact of genetic testing for HNPCC. *Am J Med Genet*. 2001;103(1):9-15.
84. Human Fertilisation and Embryology Authority S and IDW team,. PGD conditions licensed by the HFEA - testing and screening.
85. Dewanwala A, Chittenden A, Rosenblatt M, et al. Attitudes toward childbearing and prenatal testing in individuals undergoing genetic testing for Lynch Syndrome. *Fam Cancer*. 2011;10(3):549-556.
86. Barrow P, Khan M, Lalloo F, Evans DG, Hill J. Systematic review of the impact of registration and screening on colorectal cancer incidence and mortality in familial adenomatous polyposis and Lynch syndrome. *Br J Surg*. 2013;100(13):1719-1731.
87. Wilding a., Ingham SL, Lalloo F, et al. Life expectancy in hereditary cancer predisposing diseases: an observational study. *J Med Genet*. 2012;49(4):264-269.
88. Kratz CP, Achatz MI, Brugières L, et al. Cancer Screening Recommendations for Individuals with Li-Fraumeni Syndrome. *Clin Cancer Res*. 2017;23(11):e38-e45.
89. Saya S, Killick E, Thomas S, et al. Baseline results from the UK SIGNIFY study: a whole-body MRI screening study in TP53 mutation carriers and matched controls. *Fam Cancer*. 2017;16(3):433-440.
90. Ballinger ML, Best A, Mai PL, et al. Baseline Surveillance in Li-Fraumeni Syndrome Using Whole-Body Magnetic Resonance Imaging: A Meta-analysis. *JAMA Oncol*. 2017;3(12):1634-1639.
91. Menko FH, Maher ER, Schmidt LS, et al. Hereditary leiomyomatosis and renal cell cancer (HLRCC): renal cancer risk, surveillance and treatment. *Fam Cancer*. 2014:637-644.
92. Rebbeck TR, Friebel T, Lynch HT, et al. Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: the PROSE Study Group. *J Clin Oncol*. 2004;22(6):1055-1062.
93. Kauff ND, Satagopan JM, Robson ME, et al. Risk-reducing salpingo-oophorectomy in women with a BRCA1 or BRCA2 mutation. *N Engl J Med*. 2002;346(21):1609-1615.
94. Rebbeck TR, Lynch HT, Neuhausen SL, et al. Prophylactic oophorectomy in carriers of BRCA1 or BRCA2 mutations. *N Engl J Med*. 2002;346(21):1616-1622.
95. Bebbington Hatcher M, Fallowfield LJ. A qualitative study looking at the psychosocial implications of bilateral prophylactic mastectomy. *Breast*. 2003;12(1):1-9.
96. Fitzgerald RC, Hardwick R, Huntsman D, et al. Hereditary diffuse gastric cancer: Updated

consensus guidelines for clinical management and directions for future research. *J Med Genet*. 2010;47(7):436-444.

97. Jasperson KW, Patel SG, Ahnen DJ. *APC-Associated Polyposis Conditions*. (Pagon RA, Adam MP, Bird TD et al., ed.). University of Washington, Seattle; 1993.
98. Burn J, Sheth H. The role of aspirin in preventing colorectal cancer. *Br Med Bull*. 2016;119(1):17-24.
99. Burn J, Gerdes A-M, Macrae F, et al. Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. *Lancet (London, England)*. 2011;378(9809):2081-2087.
100. Vasen HFA, Blanco I, Aktan-Collan K, et al. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. *Gut*. 2013;62(6):812-823.
101. Johnson RL, Rothman AL, Xie J, et al. Human homolog of patched, a candidate gene for the basal cell nevus syndrome. *Science*. 1996;272(5268):1668-1671.
102. Hahn H, Wicking C, Zaphiropoulous PG, et al. Mutations of the human homolog of Drosophila patched in the nevoid basal cell carcinoma syndrome. *Cell*. 1996;85(6):841-851.
103. Gailani MR, Bale SJ, Leffell DJ, et al. Developmental defects in Gorlin syndrome related to a putative tumor suppressor gene on chromosome 9. *Cell*. 1992;69(1):111-117.
104. Aszterbaum M, Yauch RL, Ph D, et al. Inhibiting the Hedgehog Pathway in Patients with the Basal-Cell Nevus Syndrome. 2012.
105. National Institute for Health and Care Excellence. *Vismodegib for Treating Basal Cell Carcinoma*.; 2017.
106. Farmer H, McCabe N, Lord CJ, et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*. 2005;434(7035):917-921.
107. Billroth T. *Die Allgemeine Chirurgische Pathologie and Therapie. 51 Vorlesungen. Ein Handbuch Fur Studierende and Artze*. Berlin: G Reimer; 1889.
108. Berrino F, De Angelis R, Sant M, et al. Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995-99: results of the EUROCARE-4 study. *Lancet Oncol*. 2007;8:773-783.
109. Mariotto AB, Rowland JH, Ries L a G, Scoppa S, Feuer EJ. Multiple cancer prevalence: a growing challenge in long-term survivorship. *Cancer Epidemiol Biomarkers Prev*. 2007;16(3):566-571.
110. Boice JD, Curtis RE, Kleinerman R a, Flannery JT, Fraumeni JF. Multiple primary cancers in Connecticut, 1935-82. *Yale J Biol Med*. 1986;59(5):533-545.
111. Rosso S, De Angelis R, Ciccolallo L, et al. Multiple tumours in survival estimates. *Eur J Cancer*. 2009;45:1080-1094.
112. Travis LB. The epidemiology of second primary cancers. *Cancer Epidemiol Biomarkers Prev*.

2006;15:2020-2026.

113. Dong C, Hemminki K. Second primary neoplasms in 633,964 cancer patients in Sweden, 1958-1996. *Int J Cancer*. 2001;93:155-161.
114. Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science (80-)*. 2015;347(6217):78-81.
115. Ahmad AS, Ormiston-Smith N, Sasieni PD. Trends in the lifetime risk of developing cancer in Great Britain: Comparison of risk for those born from 1930 to 1960. *Br J Cancer*. 2015;112(5):943-947.
116. Brobeil A, Rapaport D, Wells K, et al. Multiple primary melanomas: implications for screening and follow-up programs for melanoma. *Ann Surg Oncol*. 1997;4(1):19-23.
117. Bagni O, Filippi L, Schillaci O. Incidental Detection of Colorectal Cancer Via 18F-Choline PET/CT in a Patient With Recurrent Prostate Cancer: Usefulness of Early Images. *Clin Nucl Med*. 2015;40(6):e328-30.
118. Bertagna F, Bertoli M, Treglia G, Manenti S, Salemme M, Giubbini R. Incidental 11C-choline PET/CT uptake due to esophageal carcinoma in a patient studied for prostate cancer. *Clin Nucl Med*. 2014;39(10):e442-4.
119. Moletta L, Bissoli S, Fantin A, Passuello N, Valmasoni M, Sperti C. PET/CT incidental detection of second tumor in patients investigated for pancreatic neoplasms. *BMC Cancer*. 2018;18(1):531.
120. Zaino R, Whitney C, Brady MF, DeGeest K, Burger RA, Buller RE. Simultaneously detected endometrial and ovarian carcinomas--a prospective clinicopathologic study of 74 cases: a gynecologic oncology group study. *Gynecol Oncol*. 2001;83(2):355-362.
121. Abe O, Abe R, Enomoto K, Kikuchi K. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*. 1998;351:1451-1467.
122. De Meerleer G, Khoo V, Escudier B, et al. Radiotherapy for renal-cell carcinoma. *Lancet Oncol*. 2014;15(4):e170-7.
123. Allan JM, Travis LB. Mechanisms of therapy-related carcinogenesis. *Nat Rev Cancer*. 2005;5(12):943-955.
124. Ng J, Shuryak I. Minimizing second cancer risk following radiotherapy: current perspectives. *Cancer Manag Res*. 2015;7:1-11.
125. Preston DL, Ron E, Tokuoka S, et al. Solid cancer incidence in atomic bomb survivors: 1958-1998. *Radiat Res*. 2007;168(1):1-64.
126. Nikiforov YE. Radiation-induced thyroid cancer: What we have learned from chernobyl. *Endocr Pathol*. 2006;17(4):307-317.
127. Inskid PD. Second Cancers Following Radiotherapy. In: Neugut A, Meadows A, Robinson E, eds. *Multiple Primary Cancers*. Mishawaka: Lippincott Williams and Wilkins; 1999:91-136.
128. Curtis RE, Boice JD, Stovall M, et al. Relationship of leukemia risk to radiation dose

following cancer of the uterine corpus. *J Natl Cancer Inst*. 1994;86(17):1315-1324.

129. Yahalom J, Petrek J. Breast cancer in patients irradiated for Hodgkin's disease: a clinical and pathologic analysis of 45 events in 37 patients. *J Clin …*. 1992;10(11):1674-1681.
130. Crump M, Hodgson D. Secondary breast cancer in Hodgkin's lymphoma survivors. *J Clin Oncol*. 2009;27(26):4229-4231.
131. De Bruin ML, Sparidans J, Van't Veer MB, et al. Breast cancer risk in female survivors of Hodgkin's lymphoma: Lower risk after smaller radiation volumes. *J Clin Oncol*. 2009;27(26):4239-4246.
132. Ezoe S. Secondary leukemia associated with the anti-cancer agent, etoposide, a topoisomerase II inhibitor. *Int J Environ Res Public Health*. 2012;9(7):2444-2453.
133. Travis LB, Curtis RE, Glimelius B, et al. Bladder and kidney cancer following cyclophosphamide therapy for non-Hodgkin's lymphoma. *J Natl Cancer Inst*. 1995;87(7):524-530.
134. Rubino C, Adjadj E, Guérin S, et al. Long-term risk of second malignant neoplasms after neuroblastoma in childhood: Role of treatment. *Int J Cancer*. 2003;107(5):791-796.
135. Roychoudhuri R, Evans H, Robinson D, Møller H. Radiation-induced malignancies following radiotherapy for breast cancer. *Br J Cancer*. 2004;91(5):868-872.
136. Breslow NE, Takashima JR, Whitton JA, Moksness J, D'Angio GJ, Green DM. Second malignant neoplasms following treatment for Wilms' tumor: A report from the National Wilms' Tumor Study Group. *J Clin Oncol*. 1995;13(8):1851-1859.
137. Birdwell SH, Hancock SL, Varghese A, Cox RS, Hoppe RT. Gastrointestinal cancer after treatment of Hodgkin's disease. *Int J Radiat Oncol Biol Phys*. 1997;37(1):67-73.
138. Marín-Gutzke M, Sánchez-Olaso A, Berenguer B, et al. *Basal Cell Carcinoma in Childhood after Radiation Therapy: Case Report and Review.* Annals of plastic surgery 53, 593-595 (2004).
139. Evans DGR, Birch JM, Ramsden RT, Sharif S, Baser ME. Malignant transformation and new primary tumours after therapeutic radiation for benign disease: substantial risks in certain tumour prone syndromes. *J Med Genet*. 2006;43(4):289-294.
140. Rodriguez-Antona C, Ingelman-Sundberg M. Cytochrome P450 pharmacogenetics and cancer. *Oncogene*. 2006;25(11):1679-1691.
141. Levi F, Randimbison L, Te VC, La Vecchia C. Second primary cancers in patients with lung carcinoma. *Cancer*. 1999;86(1):186-190.
142. Health and Social Care Information Centre. *Statistics on Smoking - England 2014*.; 2014.
143. Health Survey for England. *Health Survey for England - 2013*.; 2014.
144. De Leon J, Rendon DM, Baca-Garcia E, et al. Association between smoking and alcohol use in the general population: Stable and unstable odds ratios across two years in two different countries. *Alcohol Alcohol*. 2007;42(3):252-257.

145. Hemminki K, Zhang H, Sundquist J, Lorenzo Bermejo J. Modification of risk for subsequent cancer after female breast cancer by a family history of breast cancer. *Breast Cancer Res Treat*. 2008;111:165-169.
146. Hemminki K, Li X, Dong C. Second primary cancers after sporadic and familial colorectal cancer. *Cancer Epidemiol Biomarkers Prev*. 2001;10:793-798.
147. Kraemer K. Xeroderma Pigmentosum. In: Pagon RA, Adam MP, Bird TD et al., ed. *GeneReviews™ [Internet]*. Seattle: University of Washington, Seattle; 2016.
148. Evans D. Nevoid Basal Cell Carcinoma Syndrome. In: Pagon RA, Adam MP, Bird TD et al., ed. *GeneReviews™ [Internet]*. Seattle: University of Washington, Seattle; 2018.
149. Friedman J. Neurofibromatosis 1. In: Pagon RA, Adam MP, Bird TD et al., ed. *GeneReviews™ [Internet]*. University of Washington, Seattle; 2012.
150. Evans DG. Neurofibromatosis 2. In: Pagon RA, Adam MP, Bird TD et al., ed. *GeneReviews™ [Internet]*. Seattle: University of Washington, Seattle; 2018.
151. Gonzalez KD, Noltner KA, Buzin CH, et al. Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. *J Clin Oncol*. 2009;27(8):1250-1256.
152. Lim W, Olschwang S, Keller JJ, et al. Relative frequency and morphology of cancers in STK11 mutation carriers. *Gastroenterology*. 2004;126(7):1788-1794.
153. Kleinerman R, Yu C-L, Little MP, et al. Variation of second cancer risk by family history of retinoblastoma among long-term survivors. *J Clin Oncol*. 2012;30(9):950-957.
154. Metcalfe K a, Lynch HT, Ghadirian P, et al. The risk of ovarian cancer after breast cancer in BRCA1 and BRCA2 carriers. *Gynecol Oncol*. 2005;96(1):222-226.
155. Win AK, Lindor NM, Winship I, et al. Risks of colorectal and other cancers after endometrial cancer for women with Lynch syndrome. *J Natl Cancer Inst*. 2013;105(4):274-279.
156. International Agency for Research on Cancer. *International Rules for Multiple Primary Cancers (ICD-O Third Edition)*. Lyon; 2004.
157. World Health Organisation. *International Classification of Diseases for Oncology (ICD-O)*. World Health Organisation Press; 2014.
158. Illumina. TruSeq DNA PCR-Free. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_truseq_dna_pcr_free_sample_prep.pdf. Published 2013.
159. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 2007;81(3):559-575.
160. Gudbjartsson DF, Sulem P, Helgason H, et al. Sequence variants from whole genome sequencing a large group of Icelanders. *Sci data*. 2015;2:150011.
161. Lek M, Karczewski KJ, Minikel E V., et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291.
162. Jun G, Flickinger M, Hetrick KN, et al. Detecting and estimating contamination of human

DNA samples in sequencing and array-based genotype data. *Am J Hum Genet*. 2012;91(5):839-848.

163. Walter K, Min JL, Huang J, et al. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526(7571):82-89.
164. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: Versatile and scalable detection of copy number variants. *Bioinformatics*. 2016;32(15):2375-2377.
165. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222.
166. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.
167. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010;26(22):2867-2873.
168. Conomos MP, Thornton T. Title GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. 2016.
169. Illumina. TruSight Cancer Sequencing Panel. https://www.illumina.com/Documents/products/datasheets/datasheet_trusight_cancer.pdf. Published 2016.
170. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
171. DePristo M a, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.
172. Whitworth J, Smith PS, Martin JE, et al. Comprehensive Cancer-Predisposition Gene Testing in an Adult Multiple Primary Tumor Series Shows a Broad Range of Deleterious Variants and Atypical Tumor Phenotypes. *Am J Hum Genet*. 2018;103(1):3-18.
173. Wonderling D, Hopwood P, Cull A, et al. A descriptive study of UK cancer genetics services : an. *Br J Cancer*. 2001;85:166-170.
174. Pujol P, Lyonnet DS, Frebourg T, et al. Lack of referral for genetic counseling and testing in BRCA1/2 and Lynch syndromes: a nationwide study based on 240,134 consultations and 134,652 genetic tests. *Breast Cancer Res Treat*. 2013;141(1):135-144.
175. Curtis RE, Freedman DM, Ron E, Ries LAG, Hacker DG, Edwards BK, Tucker MA FJJ (eds). *New Malignancies Among Cancer Survivors: SEER Cancer Registries, 1973-2000*. Bethesda, MD: National Cancer Institute; 2000.
176. The National Cancer Registration Service - Eastern Office. http://ecric.nhs.uk/. Accessed June 4, 2015.
177. Cancer Research UK. Cancer Incidence Statistics. http://www.cancerresearchuk.org/health-

professional/cancer-statistics/incidence. Accessed April 8, 2016.

178. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.r-project.org/. Published 2017.
179. Carss KJ, Arno G, Erwood M, et al. Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. *Am J Hum Genet*. 2017;100(1):75-90.
180. Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci*. 2015;112(17):5473-5478.
181. Suzuki T, Tsurusaki Y, Nakashima M, et al. Precise detection of chromosomal translocation or inversion breakpoints by whole-genome sequencing. *J Hum Genet*. 2014;59(12):649-654.
182. Jafri M, Wake NC, Ascher DB, et al. Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov*. 2015;5(7):723-729.
183. Farrell CM, O'Leary NA, Harte RA, et al. Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res*. 2014;42(Database issue):D865-72.
184. Flicek P, Ahmed I, Amode MR, et al. Ensembl 2013. *Nucleic Acids Res*. 2013;41(Database issue):D48-55.
185. Smedley D, Haider S, Durinck S, et al. The BioMart community portal: An innovative alternative to large, centralized data repositories. *Nucleic Acids Res*. 2015;43(W1):W589-W598.
186. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
187. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2015;44(D1):D862-8.
188. Stenson PD, Ball E V., Mort M, et al. Human Gene Mutation Database (HGMD®): 2003 Update. *Hum Mutat*. 2003;21(6):577-581.
189. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315.
190. University Medical Center Groningen. FaCD Online. http://www.familialcancerdatabase.nl/default.aspx. Accessed February 1, 2017.
191. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.
192. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424.
193. Finn RD, Coggill P, Eberhardt RY, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2015;44(D1):D279-D285.

194. James Kent W, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006.
195. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet*. 2015;16(3):172-183.
196. Whitworth J, Hoffman J, Chapman C, et al. A clinical and genetic analysis of multiple primary cancer referrals to genetics services. *Eur J Hum Genet*. 2014;23(February):1-7.
197. Thompson D, Duedal S, Kirner J, et al. Cancer risks and mortality in heterozygous ATM mutation carriers. *J Natl Cancer Inst*. 2005;97(11):813-822.
198. Weischer M, Bojesen SE, Tybjaerg-Hansen A, Axelsson CK, Nordestgaard BG. Increased risk of breast cancer associated with CHEK2*1100delC. *J Clin Oncol*. 2007;25(1):57-63.
199. Lalloo F, Varley J, Ellis D, et al. Prediction of pathogenic mutations in patients with early-onset breast cancer by family history. *Lancet*. 2003;361(9363):1101-1102.
200. Schrader KA, Cheng DT, Joseph V, et al. Germline variants in targeted tumor sequencing using matched normal DNA. *JAMA Oncol*. 2016;2(1):104-111.
201. Mandelker D, Zhang L, Kemel Y, et al. Mutation detection in patients with advanced cancer by universal sequencing of cancer-related genes in tumor and normal DNA vs guideline-based germline testing. *JAMA - J Am Med Assoc*. 2017;318(9):825-835.
202. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med*. 2015;373(24):2336-2346.
203. Parsons DW, Roy A, Yang Y, et al. Diagnostic yield of clinical tumor and germline whole-exome sequencing for children with solid tumors. *JAMA Oncol*. 2016;2(5):616-624.
204. Mody RJ, Wu YM, Lonigro RJ, et al. Integrative clinical sequencing in the management of refractory or relapsed cancer in youth. *JAMA - J Am Med Assoc*. 2015;314(9):913-925.
205. Pearlman R, Frankel WL, Swanson B, et al. Prevalence and Spectrum of Germline Cancer Susceptibility Gene Mutations Among Patients With Early-Onset Colorectal Cancer. *JAMA Oncol*. 2016;354(26):2751-2763.
206. Cybulski C, Górski B, Huzarski T, et al. CHEK2 is a multiorgan cancer susceptibility gene. *Am J Hum Genet*. 2004;75(6):1131-1135.
207. Aoude LG, Xu M, Zhao ZZ, et al. Assessment of PALB2 as a candidate melanoma susceptibility gene. *PLoS One*. 2014;9(6):e100683.
208. Tomlinson IPM, Alam NA, Rowan AJ, et al. Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. *Nat Genet*. 2002;30(april):406-410.
209. Castro-Vega LJ, Buffet A, De Cubas AA, et al. Germline mutations in FH confer predisposition to malignant pheochromocytomas and paragangliomas. *Hum Mol Genet*. 2014;23(9):2440-2446.
210. Clark GR, Sciacovelli M, Gaude E, et al. Germline FH mutations presenting with

pheochromocytoma. *J Clin Endocrinol Metab*. 2014;99(10):E2046-E2050.

211. Coughlin EM, Christensen E, Kunz PL, et al. Molecular analysis and prenatal diagnosis of human fumarase deficiency. *Mol Genet Metab*. 1998;63(4):254-262.
212. Whitworth J, Skytte A-B, Sunde L, et al. Multilocus inherited neoplasia alleles syndrome: A case series and review. *JAMA Oncol*. 2015;2(3):373-379.
213. National Human Genome Research Institute. The Cost of Sequencing a Human Genome. https://www.genome.gov/sequencingcosts/. Published 2016.
214. Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol*. 2015;3:92.
215. Feero WG. Clinical application of whole-genome sequencing: Proceed with care. *JAMA - J Am Med Assoc*. 2014;311(10):1017-1019.
216. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
217. Ipe J, Swart M, Burgess KS, Skaar TC. High-throughput assays to assess the functional impact of genetic variants: A road towards genomic-driven medicine. *Clin Transl Sci*. 2017;10(2):67-77.
218. Evans DGR, Lalloo F, Wallace A, Rahman N. Update on the Manchester Scoring System for BRCA1 and BRCA2 testing. *J Med Genet*. 2005;42(7):e39.
219. Cancer Research UK. Cancer incidence for common cancers. http://www.cancerresearchuk.org/cancer-info/cancerstats/incidence/commoncancers/. Published 2013. Accessed November 1, 2013.
220. Rare Cancers Europe. About Rare Cancers. http://www.rarecancerseurope.org/About-Rare-Cancers. Published 2017. Accessed October 10, 2017.
221. Cancer Research UK. Statistics on Preventable Cancers. http://www.cancerresearchuk.org/health-professional/cancer-statistics/risk/preventable-cancers.
222. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
223. Strachan T, Read A. *Human Molecular Genetics*. Third. New York: Garland Publishing; 2004.
224. Ju YS, Martincorena I, Gerstung M, et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*. 2017;543(7647):714-718.
225. Moyhuddin a, Baser ME, Watson C, et al. Somatic mosaicism in neurofibromatosis 2: prevalence and risk of disease transmission to offspring. *J Med Genet*. 2003;40(6):459-463.
226. Renaux-Petel M, Charbonnier F, Théry J-C, et al. Contribution of de novo and mosaic TP53 mutations to Li-Fraumeni syndrome. *J Med Genet*. 2018;55(3):173-180.
227. Delon I, Taylor A, Molenda A, et al. A germline mosaic BRCA1 exon deletion in a woman with bilateral basal-like breast cancer. *Clin Genet*. 2013;84(3):297-299.

228. Zhuang Z, Yang C, Lorenzo F, et al. Somatic HIF2A gain-of-function mutations in paraganglioma with polycythemia. *N Engl J Med*. 2012;367(10):922-930.
229. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
230. National Center for Biotechnology Information. *The NCBI Handbook*. Bethesda, MD; 2002.
231. Barone G, Groom A, Reiman A, Srinivasan V, Byrd PJ, Taylor AMR. Modeling ATM mutant proteins from missense changes confirms retained kinase activity. *Hum Mutat*. 2009;30(8):1222-1230.
232. Bhatia S, Sklar C. Second cancers in survivors of childhood cancer. *Nat Rev Cancer*. 2002;2(2):124-132.
233. Leung W, Ribeiro RC, Hudson M, et al. Second malignancy after treatment of childhood acute myeloid leukemia. *Leukemia*. 2001;15(1):41-45.
234. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*. 2012;481(7382):506-510.
235. Evans DGR, Ramsden RT, Shenton A, et al. Mosaicism in neurofibromatosis type 2: An update of risk based on uni/bilaterality of vestibular schwannoma at presentation and sensitive mutation analysis including multiple ligation-dependent probe amplification. *J Med Genet*. 2007;44(7):424-428.
236. Mani N, Slevin N, Hudson A. What Three Wise Men have to say about diagnosis. *BMJ*. 2011;343(dec19 2):d7769-d7769.
237. Eurodis. What is a rare disease? http://www.eurordis.org/content/what-rare-disease. Published 2015. Accessed May 14, 2015.
238. Bakry D, Aronson M, Durno C, et al. Genetic and clinical determinants of constitutional mismatch repair deficiency syndrome: Report from the constitutional mismatch repair deficiency consortium. *Eur J Cancer*. 2014;50(5):987-996.
239. Augustyn AM, Agostino NM, Namey TL, Nair S, Martino M a. Two patients with germline mutations in both BRCA1 and BRCA2 discovered unintentionally: a case series and discussion of BRCA testing modalities. *Breast Cancer Res Treat*. 2011;129(2):629-634.
240. Caldes T. A breast cancer family from Spain with germline mutations in both the BRCA1 and BRCA2 genes. *J Med Genet*. 2002;39(8):44e-44.
241. Choi DH, Lee MH, Haffty BG. Double heterozygotes for non-Caucasian families with mutations in BRCA-1 and BRCA-2 genes. *Breast J*. 2006;12(3):216-220.
242. Heidemann S, Fischer C, Engel C, et al. Double heterozygosity for mutations in BRCA1 and BRCA2 in German breast cancer patients: implications on test strategies and clinical management. *Breast Cancer Res Treat*. 2012;134(3):1229-1239.
243. Leegte B, van der Hout a H, Deffenbaugh a M, et al. Phenotypic expression of double heterozygosity for BRCA1 and BRCA2 germline mutations. *J Med Genet*. 2005;42(3):e20.

244. Liede a, Rehal P, Vesprini D, Jack E, Abrahamson J, Narod S a. A breast cancer patient of Scottish descent with germ-line mutations in BRCA1 and BRCA2. *Am J Hum Genet*. 1998;62(6):1543-1544.
245. Loubser F, de Villiers JNP, van der Merwe NC. Two double heterozygotes in a South African Afrikaner family: implications for BRCA1 and BRCA2 predictive testing. *Clin Genet*. 2012;82(6):599-600.
246. Moslehi R, Russo D, Phelan C, Jack E, Antman K, Narod S. An unaffected individual from a breast / ovarian cancer family with germline mutations in both BRCA 1 and BRCA 2. *Clin Genet*. 2000;57:70-73.
247. Musolino A, Naldi N, Michiara M, et al. A breast cancer patient from Italy with germline mutations in both the BRCA1 and BRCA2 genes. *Breast Cancer Res Treat*. 2005;91(2):203-205.
248. Noh JM, Choi DH, Nam SJ, et al. Characteristics of double heterozygosity for BRCA1 and BRCA2 germline mutations in Korean breast cancer patients. *Breast Cancer Res Treat*. 2012;131(1):217-222.
249. Pilato B, De Summa S, Danza K, Lambo R, Paradiso A, Tommasi S. Maternal and paternal lineage double heterozygosity alteration in familial breast cancer: a first case report. *Breast Cancer Res Treat*. 2010;124(3):875-878.
250. Smith M, Fawcett S, Sigalas E, et al. Familial breast cancer: double heterozygosity for BRCA1 and BRCA2 mutations with differing phenotypes. *Fam Cancer*. 2008;7(2):119-124.
251. Tesoriero A, Andersen C, Southey M, et al. De Novo BRCA1 Mutation in a Patient with Breast Cancer and an inherited BRCA2 Mutation. *Am J Hum Genet*. 1999;65:567-569.
252. Zuradelli M, Peissel B, Manoukian S, et al. Four new cases of double heterozygosity for BRCA1 and BRCA2 gene mutations: clinical, pathological, and family characteristics. *Breast Cancer Res Treat*. 2010;124(1):251-258.
253. Ramus SJ, Friedman LS, Gayther SA, et al. A breast/ovarian cancer patient with germline mutations in both BRCA1 and BRCA2. *Nat Genet*. 1997;15(1):14-15.
254. Randall TC, Bell KA, Rebane BA, Rubin SC, Boyd J, Ph D. CASE REPORT Germline Mutations of the BRCA1 and BRCA2 Genes in a Breast and Ovarian Cancer Patient 1. *Gynecol Oncol*. 1998;70:432-434.
255. Bell DW, Erban J, Sgroi DC, Haber DA. Selective Loss of Heterozygosity in Multiple Breast Cancers from a Carrier of Mutations in Both BRCA1 and BRCA2 Advances in Brief Selective Loss of Heterozygosity in Multiple Breast Cancers from a Carrier of Mutations in Both BRCA1 and BRCA2 1. *Cancer Res*. 2002;62:2741-2743.
256. Friedman E, Bar-Sade Bruchim R, Kruglikova A, et al. Double heterozygotes for the Ashkenazi founder mutations in BRCA1 and BRCA2 genes. *Am J Hum Genet*. 1998;63(4):1224-1227.

257. Frank TS, Deffenbaugh AM, Reid JE, et al. Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: Analysis of 10,000 individuals. *J Clin Oncol*. 2002;20(6):1480-1490.
258. Toro JR, Wei M-H, Glenn GM, et al. BHD mutations, clinical and molecular genetic investigations of Birt-Hogg-Dubé syndrome: a new series of 50 families and a review of published reports. *J Med Genet*. 2008;45(6):321-331.
259. Schmidt LS, Nickerson ML, Warren MB, et al. Germline BHD-mutation spectrum and phenotype analysis of a large cohort of families with Birt-Hogg-Dubé syndrome. *Am J Hum Genet*. 2005;76(6):1023-1033.
260. Orphanet. Prevalence of rare diseases: Bibliographic data. http://www.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_alphabetical_list.pdf. Published 2014. Accessed May 14, 2015.
261. Toro JR. Birt-Hogg-Dubé Syndrome. In: Pagon RA, Adam MP, Bird TD et al., ed. *GeneReviews™ [Internet]*. University of Washington, Seattle; 2008.
262. Goddard NC, McIntyre A, Summersgill B, Gilbert D, Kitazawa S, Shipley J. KIT and RAS signalling pathways in testicular germ cell tumours: new data and a review of the literature. *Int J Androl*. 2007;30(4):337-48; discussion 349.
263. Wang Y, Gan Y, Tan Z, et al. TDRG1 functions in testicular seminoma are dependent on the PI3K/Akt/mTOR signaling pathway. *Onco Targets Ther*. 2016;9:409-420.
264. Tidyman WE, Rauen KA. The RASopathies: developmental syndromes of Ras/MAPK pathway dysregulation. *Curr Opin Genet Dev*. 2009;19(3):230-236.
265. Tsun Z-Y, Bar-Peled L, Chantranupong L, et al. The folliculin tumor suppressor is a GAP for the RagC/D GTPases that signal amino acid levels to mTORC1. *Mol Cell*. 2013;52(4):495-505.
266. Mannuel HD, Mitikiri N, Khan M, Hussain A. Testicular germ cell tumors. *Curr Opin Oncol*. 2012;24(3):266-271.
267. Woodward ER, Ricketts C, Killick P, et al. Familial non-VHL clear cell (conventional) renal cell carcinoma: Clinical features, segregation analysis, and mutation analysis of FLCN. *Clin Cancer Res*. 2008;14(18):5925-5930.
268. Petitjean A, Mathe E, Kato S, et al. Impact of Mutant p53 Functional Properties on TP53 Mutation Patterns and Tumor Phenotype : Lessons from Recent Developments in the IARC TP53 Database. *Hum Mutat*. 2007;28(February):622-629.
269. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009;4(7):1073-1081.
270. Adzhubei I a, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-249.
271. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations:

application to cancer genomics. *Nucleic Acids Res*. 2011;39(17):e118.

272. Nahorski MS, Lim DHK, Martin L, et al. Investigation of the Birt-Hogg-Dube tumour suppressor gene (FLCN) in familial and sporadic colorectal cancer. *J Med Genet*. 2010;47(6):385-390.
273. Eaden J a, Abrams KR, Mayberry JF. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut*. 2001;48(4):526-535.
274. Wong P, Verselis SJ, Garber JE, et al. Prevalence of early onset colorectal cancer in 397 patients with classic Li-Fraumeni syndrome. *Gastroenterology*. 2006;130(1):73-79.
275. Kohlmann W, Gruber SB. *Lynch Syndrome*. (Pagon RA, Adam MP, Bird TD et al., ed.). University of Washington, Seattle; 2014.
276. Thompson BA, Spurdle AB, Plazzer J-P, et al. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat Genet*. 2014;46(2):107-115.
277. Frayling IM, van Minkelen R, Baralle D, et al. Predisposition to breast cancer in NF1 occurs before, but not after, age 50y and is unrelated to NF1 gene mutation type or site (Poster PM12.210). In: *European Society of Human Genetics*. Glasgow, UK; 2015.
278. Sharif S, Moran A, Huson SM, et al. Women with neurofibromatosis 1 are at a moderately increased risk of developing breast cancer and should be considered for early screening. *J Med Genet*. 2007;44(8):481-484.
279. Burgoyne AM, Somaiah N, Sicklick JK. Gastrointestinal stromal tumors in the setting of multiple tumor syndromes. *Curr Opin Oncol*. 2014;26(4):408-414.
280. Postow MA, Robson ME. Inherited gastrointestinal stromal tumor syndromes: mutations, clinical features, and therapeutic implications. *Clin Sarcoma Res*. 2012;2(1):16.
281. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415-421.
282. Polak P, Kim J, Braunstein LZ, et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat Genet*. 2017;49(10):1476-1486.
283. Connor AA, Denroche RE, Jang GH, et al. Association of Distinct Mutational Signatures With Correlates of Increased Immune Activity in Pancreatic Ductal Adenocarcinoma. *JAMA Oncol*. 2017;3(6):774-783.
284. Sulkowski PL, Sundaram RK, Oeck S, et al. Krebs-cycle-deficient hereditary cancer syndromes are defined by defects in homologous-recombination DNA repair. *Nat Genet*. 2018;50(8):1086-1092.
285. Casey RT, Warren AY, Martin JE, et al. Clinical and Molecular Features of Renal and Pheochromocytoma/Paraganglioma Tumor Association Syndrome (RAPTAS): Case Series and Literature Review. *J Clin Endocrinol Metab*. 2017;102(11):4013-4022.

286. Comino-Méndez I, Gracia-aznárez FJ, Schiavi F, et al. Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nat Genet*. 2011;43(7):663-667.
287. Borg A, Isola J, Chen J, et al. Germline BRCA1 and HMLH1 Mutations in a Family with Male and Female Breast Carcinoma. *Int J Cancer*. 2000;85:796-800.
288. Pedroni M, Di Gregorio C, Cortesi L, et al. Double heterozygosity for BRCA1 and hMLH1 gene mutations in a 46-year-old woman with five primary tumors. *Tech Coloproctol*. 2014;18(3):285-289.
289. Kast K, Neuhann TM, Görgens H, et al. Germline truncating-mutations in BRCA1 and MSH6 in a patient with early onset endometrial cancer. *BMC Cancer*. 2012;12:531.
290. Thiffault I, Hamel N, Pal T, et al. Germline truncating mutations in both MSH2 and BRCA2 in a single kindred. *Br J Cancer*. 2004;90(2):483-491.
291. Foppiani L, Forzano F, Ceccherini I, et al. Uncommon association of germline mutations of RET proto-oncogene and CDKN2A gene. *Eur J Endocrinol*. 2008;158(3):417-422.
292. Monnerat C, Chompret A, Kannengiesser C, et al. BRCA1, BRCA2, TP53, and CDKN2A germline mutations in patients with breast cancer and cutaneous melanoma. *Fam Cancer*. 2007;6(4):453-461.
293. Ghataorhe P, Kurian AW, Pickart A, et al. A carrier of both MEN1 and BRCA2 mutations: case report and review of the literature. *Cancer Genet Cytogenet*. 2007;179(2):89-92.
294. Kashiwada T, Shimizu H, Tamura K, Seyama K, Horie Y, Mizoo A. Birt-Hogg-Dube Syndrome and Familial Adenomatous Polyposis: An Association or a Coincidence? *Intern Med*. 2012;51(13):1789-1792.
295. Kilmartin DJ, Mooney DJ, Acheson RW, Payne SJ, Maher ER, Eustace P. von Hippel-Lindau disease and familial polyposis coli in the same family. *Arch Ophthalmol*. 1996;114(10):1294.
296. Mastroianno S, Torlontano M, Scillitani A, et al. Coexistence of multiple endocrine neoplasia type 1 and type 2 in a large Italian family. *Endocrine*. 2011;40(3):481-485.
297. Plon SE, Pirics ML, Nuchtern J, et al. Multiple tumors in a child with germ-line mutations in TP53 and PTEN. *N Engl J Med*. 2008;359(5):537-539.
298. Zbuk KM, Patocs A, Shealy A, Sylvester H, Miesfeldt S, Eng C. Germline mutations in PTEN and SDHC in a woman with epithelial thyroid cancer and carotid paraganglioma. *Nat Clin Pract Oncol*. 2007;4(10):608-612.
299. Lindor NM, Smyrk TC, Buehler S, et al. Multiple jejunal cancers resulting from combination of germline APC and MLH1 mutations. *Fam Cancer*. 2012;11(4):667-669.
300. Soravia C, DeLozier CD, Dobbie Z, et al. Double frameshift mutations in APC and MSH2 in the same individual. *Int J Colorectal Dis*. 2005;20(5):466-470.
301. Uhrhammer N, Bignon Y-J. Report of a family segregating mutations in both the APC and MSH2 genes: juvenile onset of colorectal cancer in a double heterozygote. *Int J Colorectal Dis*. 2008;23(11):1131-1135.

302. Scheenstra R, Rijcken FEM, Koornstra JJ, et al. Rapidly progressive adenomatous polyposis in a patient with germline mutations in both the APC and MLH1 genes: the worst of two worlds. *Gut*. 2003;52:898-899.
303. van Puijenbroek M, Nielsen M, Reinards THCM, et al. The natural history of a combined defect in MSH6 and MUTYH in a HNPCC family. *Fam Cancer*. 2007;6(1):43-51.
304. Pern F, Bogdanova N, Schürmann P, et al. Mutation analysis of BRCA1, BRCA2, PALB2 and BRD7 in a hospital-based series of German patients with triple-negative breast cancer. *PLoS One*. 2012;7(10):e47993.
305. Ercolino T, Lai R, Giachè V, et al. Patient affected by neurofibromatosis type 1 and thyroid C-cell hyperplasia harboring pathogenic germ-line mutations in both NF1 and RET genes. *Gene*. 2014;536(2):332-335.
306. Campos B, Balmaña J, Gardenyes J, et al. Germline mutations in NF1 and BRCA1 in a family with neurofibromatosis type 1 and early-onset breast cancer. *Breast Cancer Res Treat*. 2013;139(2):597-602.
307. Bell K, Hodgson N, Levine M, Sadikovic B, Zbuk K. Double heterozygosity for germline mutations in BRCA1 and p53 in a woman with early onset breast cancer. *Breast Cancer Res Treat*. 2014;146(2):447-450.
308. Ahlborn LB, Steffensen AY, Jønson L, et al. Identification of a breast cancer family double heterozygote for RAD51C and BRCA2 gene mutations. *Fam Cancer*. 2015;14(1):129-133.
309. Crawford B, Adams SB, Sittler T, et al. Multi-gene panel testing for hereditary cancer predisposition in unsolved high-risk breast and ovarian cancer patients. *Breast Cancer Res Treat*. 2017;163(2):383-390.
310. Eliade M, Skrzypski J, Baurand A, et al. The transfer of multigene panel testing for hereditary breast and ovarian cancer to healthcare: What are the implications for the management of patients and families? *Oncotarget*. 2017;8(2):1957-1971.
311. Francies FZ, Wainstein T, De Leeneer K, et al. BRCA1, BRCA2 and PALB2 mutations and CHEK2 c.1100delC in different South African ethnic groups diagnosed with premenopausal and/or triple negative breast cancer. *BMC Cancer*. 2015;15(1):912.
312. Goehringer C, Sutter C, Kloor M, et al. Double germline mutations in APC and BRCA2 in an individual with a pancreatic tumor. *Fam Cancer*. 2017;16(2):303-309.
313. Meynard G, Mansi L, Lebahar P, et al. First description of a double heterozygosity for BRCA1 and BRCA2 pathogenic variants in a French metastatic breast cancer patient: A case report. *Oncol Rep*. 2017;37(3):1573-1578.
314. Njoroge SW, Burgess KR, Cobleigh MA, Alnajar HH, Gattuso P, Usha L. Hereditary diffuse gastric cancer and lynch syndromes in a BRCA1/2 negative breast cancer patient. *Breast Cancer Res Treat*. 2017;166(1):315-319.
315. Nomizu T, Matsuzaki M, Katagata N, et al. A case of familial breast cancer with double

heterozygosity for BRCA1 and BRCA2 genes. *Breast Cancer*. 2015;22(5):557-561.

316. Silva-Smith R, Sussman DA. Co-occurrence of Lynch syndrome and juvenile polyposis syndrome confirmed by multigene panel testing. *Fam Cancer*. 2018;17(1):87-90.
317. Sokolenko AP, Bogdanova N, Kluzniak W, et al. Double heterozygotes among breast cancer patients analyzed for BRCA1, CHEK2, ATM, NBN/NBS1, and BLM germ-line mutations. *Breast Cancer Res Treat*. 2014;145(2):553-562.
318. Vietri MT, Molinari AM, Caliendo G, et al. Double heterozygosity in the BRCA1 and BRCA2 genes in Italian family. *Clin Chem Lab Med*. 2013;51(12):2319-2324.
319. Valle L, Rodriguez-Lopez R, Robledo M, Benetiz J, Urioste M. Concurrence of Germline Mutations in the APC and PTEN Genes in a Colonic Polyposis Family Member. *J Clin Oncol*. 2004;22(11):2252-2253.
320. Bowman-Colin C, Xia B, Bunting S, et al. Palb2 synergizes with Trp53 to suppress mammary tumor formation in a model of inherited breast cancer. *Proc Natl Acad Sci U S A*. 2013;110(21):8632-8637.
321. Jonkers J, Meuwissen R, van der Gulden H, Peterse H, van der Valk M, Berns a. Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat Genet*. 2001;29(4):418-425.
322. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57.
323. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13.
324. Baker KE, Parker R. Nonsense-mediated mRNA decay: Terminating erroneous gene expression. *Curr Opin Cell Biol*. 2004;16(3):293-299.
325. MacArthur D, Balasubramanian S, Frankish A. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science (80-)*. 2012;335(6070):1-14.
326. Saleheen D, Natarajan P, Armean IM, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*. 2017;544(7649):235-239.
327. Thompson ER, Gorringe KL, Rowley SM, et al. Reevaluation of the BRCA2 truncating allele c.9976A > T (p.Lys3326Ter) in a familial breast cancer context. *Sci Rep*. 2015;5:14800.
328. Makarov EM, Shtam TA, Kovalev RA, Pantina RA, Varfolomeeva EY, Filatov M V. The rare nonsense mutation in p53 triggers alternative splicing to produce a protein capable of inducing apoptosis. *PLoS One*. 2017;12(9):e0185126.
329. Disset A, Bourgeois CF, Benmalek N, Claustres M, Stevenin J, Tuffery-Giraud S. An exon skipping-associated nonsense mutation in the dystrophin gene uncovers a complex interplay between multiple antagonistic splicing elements. *Hum Mol Genet*. 2006;15(6):999-1013.
330. Aartsma-Rus A, Straub V, Hemmings R, et al. Development of Exon Skipping Therapies for Duchenne Muscular Dystrophy: A Critical Review and a Perspective on the Outstanding

Issues. *Nucleic Acid Ther*. 2017;27(5):251-259.

331. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med*. 2006;354(12):1264-1272.
332. Fousteri M, Mullenders LHF. Transcription-coupled nucleotide excision repair in mammalian cells: Molecular mechanisms and biological effects. *Cell Res*. 2008;18(1):73-84.
333. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499(7457):214-218.
334. Cancer Genome Atlas Research Network, Ley TJ, Miller C, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.
335. Lohr JG, Stojanov P, Carter SL, et al. Widespread Genetic Heterogeneity in Multiple Myeloma: Implications for Targeted Therapy. *Cancer Cell*. 2014;25(1):91-101.
336. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337.
337. Song Y, Li L, Ou Y, et al. Identification of genomic alterations in oesophageal squamous cell cancer. *Nature*. 2014;509(7498):91-95.
338. Lin D-C, Hao J-J, Nagata Y, et al. Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat Genet*. 2014;46(5):467-473.
339. Dulak AM, Stojanov P, Peng S, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet*. 2013;45(5):478-486.
340. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609-615.
341. Dees ND, Zhang Q, Kandoth C, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Res*. 2012;22(8):1589-1598.
342. Fadista J, Oskolkov N, Hansson O, Groop L. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics*. 2017;33(4):471-474.
343. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res*. 2013;41(Web Server issue):W77-83.
344. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29.
345. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*. 2017;45(D1):D331-D338.
346. Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010;38(Web Server issue):W214-20.
347. LOFTEE (Loss-Of-Function Transcript Effect Estimator). 2014.
348. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape

of oncogenic signatures across human cancers. *Nat Genet*. 2013;45(10):1127-1133.

349. Ruark E, Münz M, Renwick A, et al. The ICR1000 UK exome series : a resource of gene variation in an outbred population [version 1 ; referees : 1 approved] Referee Status : 2015;883:1-10.
350. Fisher RA. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *J R Stat Soc*. 1922;85(1):87.
351. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. 1995;57(1):289-300.
352. Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res*. 2017;45(D1):D619-D625.
353. Maciejowski J, De Lange T. Telomeres in cancer: Tumour suppression and genome instability. *Nat Rev Mol Cell Biol*. 2017;18(3):175-186.
354. Müezzinler A, Zaineddin AK, Brenner H. A systematic review of leukocyte telomere length and age in adults. *Ageing Res Rev*. 2013;12(2):509-519.
355. Meeker AK, Hicks JL, Iacobuzio-Donahue CA, et al. Telomere length abnormalities occur early in the initiation of epithelial carcinogenesis. *Clin Cancer Res*. 2004;10(10):3317-3326.
356. Takagi S, Kinouchi Y, Hiwatashi N, et al. Telomere shortening and the clinicopathologic characteristics of human colorectal carcinomas. *Cancer*. 1999;86(8):1431-1436.
357. Rudolph KL, Millard M, Bosenberg MW, DePinho RA. Telomere dysfunction and evolution of intestinal carcinoma in mice and humans. *Nat Genet*. 2001;28(2):155-159.
358. Savage S. Dyskeratosis Congenita. In: Adam MP, Ardinger HH, Pagon RA et al, ed. *GeneReviews® [Internet]*. Seattle: University of Washington, Seattle; 2016.
359. Talbert J. Pulmonary Fibrosis, Familial. In: Adam MP, Ardinger HH, Pagon RA et al, ed. *GeneReviews® [Internet]*. Seattle: University of Washington, Seattle; 2015.
360. Martinez-Delgado B, Yanowsky K, Inglada-Perez L, et al. Genetic anticipation is associated with Telomere shortening in hereditary breast cancer. *PLoS Genet*. 2011;7(7):e1002182.
361. Kim NW, Piatyszek MA, Prowse KR, et al. Specific association of human telomerase activity with immortal cells and cancer. *Science (80-)*. 1994;266(5193):2011-2015.
362. Horn S, Figl A, Rachakonda PS, et al. TERT promoter mutations in familial and sporadic melanoma. *Science (80-)*. 2013;339(6122):959-961.
363. Robles-Espinoza CD, Harland M, Ramsay AJ, et al. POT1 loss-of-function variants predispose to familial melanoma. *Nat Genet*. 2014;46(5):478-481.
364. Rode L, Nordestgaard BG, Bojesen SE. Long telomeres and cancer risk among 95 568 individuals from the general population. *Int J Epidemiol*. 2016;45(5):1634-1643.
365. Ramsay AJ, Quesada V, Foronda M, et al. POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet*. 2013;45(5):526-530.
366. Farmery JHR, Smith ML, NIHR BioResource - Rare Diseases, Lynch AG. Telomerecat: A

ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci Rep*. 2018;8(1):1300.

367. Lynch SM, Peek MK, Mitra N, et al. Race,ethnicity, psychosocial factors, and telomere length in a multicenter setting. *PLoS One*. 2016;11(1):e0146723.
368. Diez Roux A V., Ranjit N, Jenny NS, et al. Race/ethnicity and telomere length in the Multi-Ethnic Study of Atherosclerosis. *Aging Cell*. 2009;8(3):251-257.
369. Gardner M, Bann D, Wiley L, et al. Gender and telomere length: Systematic review and meta-analysis. *Exp Gerontol*. 2014;51(1):15-27.
370. Benitez-Buelga C, Sanchez-Barroso L, Gallardo M, et al. Impact of chemotherapy on telomere length in sporadic and familial breast cancer patients. *Breast Cancer Res Treat*. 2015;149(2):385-394.
371. Bolzán A. Effect of chemotherapeutic drugs on telomere length and telomerase activity. *Telomere and Telomerase*. 2016;3.
372. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics*. 2009;25(22):3045-3046.
373. Ma M, Ru Y, Chuang L-S, et al. Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics*. 2015;16 Suppl 8(8):S3.
374. Liu L, Dilworth D, Gao L, et al. Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat Genet*. 1999;21(1):128-132.
375. Aguet F, Brown AA, Castel SE, et al. Genetic effects on gene expression across human tissues. *Nature*. 2017;550(7675):204-213.
376. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-)*. 2007;316(5830):1497-1502.
377. Keene MA, Corces V, Lowenhaupt K, Elgin SC. DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. *Proc Natl Acad Sci*. 1981;78(1):143-146.
378. Dimitrieva S, Bucher P. UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res*. 2013;41(Database issue):D101-9.
379. Andersson R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*. 2015;37(3):314-323.
380. Safran M, Dalah I, Alexander J, et al. GeneCards Version 3: the human gene integrator. *Database (Oxford)*. 2010;2010.
381. Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*. 2017;2017(1).
382. Lizio M, Harshbarger J, Shimoji H, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol*. 2015;16(1):22.
383. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: Computational tools for

comparative genomics. *Nucleic Acids Res*. 2004;32(WEB SERVER ISS.).

384. Zhang W, Bojorquez-Gomez A, Velez DO, et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat Genet*. 2018;50(4):613-620.
385. Bell DW, Varley JM, Szydlo TE, et al. Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. *Science (80-)*. 1999;286(5449):2528-2531.
386. Bernstein JL, Teraoka SN, John EM, et al. The CHEK2*1100delC allelic variant and risk of breast cancer: screening results from the Breast Cancer Family Registry. *Cancer Epidemiol Biomarkers Prev*. 2006;15(2):348-352.
387. Gara SK, Jia L, Merino MJ, et al. Germline HABP2 Mutation Causing Familial Nonmedullary Thyroid Cancer. *N Engl J Med*. 2015;373(5):448-455.
388. Koboldt DC, Fulton RS, McLellan MD, et al. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61-70.
389. Fishbein L, Leshchiner I, Walter V, et al. Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma. *Cancer Cell*. 2017;31(2):181-193.
390. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-D745.
391. Muzny DM, Bainbridge MN, Chang K, et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337.
392. Lawrence MS, Sougnez C, Lichtenstein L, et al. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015;517(7536):576-582.
393. Bell D, Berchuck A, Birrer M, et al. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609-615.
394. Cancer Genome Atlas Research Network, Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, Davis C, Wheeler DA, Murray BA, Schmidt L, Vocke CD, Peto M, Al Mamun AA, Shinbrot E, Sethi A, Brooks S, Rathmell WK, Brooks AN, Hoadley KA, Robertson AG, Br ZR. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med*. 2016;374(2):135-145.
395. Creighton CJ, Morgan M, Gunaratne PH, et al. Comprehensivemolecular characterization of clear cell renal cell carcinoma. *Nature*. 2013;499(7456):43-49.
396. Abeshouse A, Ahn J, Akbani R, et al. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*. 2015;163(4):1011-1025.
397. Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell*. 2017;171(4):950-965.e28.
398. Guilford P, Hopkins J, Harraway J, et al. E-cadherin germline mutations in familial gastric cancer. *Nature*. 1998;392(6674):402-405.
399. Kausalya P.J., Phua C.Y HW. Association of ARVCF with Zonula occudens (ZO-1) and ZO-

2; Binding to PDZ-Domain Proteins and Cell-Cell Adhesion regulate plasma membrane and nuclear localization of ARVCF. *Mol Biol Cell*. 2004;15(12):5503-5515.

400. Lampe A, Flanigan K, Bushby K, Hicks D. Collagen Type VI-Related Disorders. In: Pagon RA, Adam MP, Bird TD et al., ed. *GeneReviews™ [Internet]*. Seattle: University of Washington, Seattle; 2012.
401. Backenroth D, He Z, Kiryluk K, et al. FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. *Am J Hum Genet*. 2018;102(5):920-942.
402. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet*. 2013;45(4):353-361.
403. Hahn MM, de Voer RM, Hoogerbrugge N, Ligtenberg MJL, Kuiper RP, van Kessel AG. The genetic heterogeneity of colorectal cancer predisposition - guidelines for gene discovery. *Cell Oncol*. 2016;39(6):491-510.
404. Jiao S, Peters U, Berndt S, et al. Estimating the heritability of colorectal cancer. *Hum Mol Genet*. 2014;23(14):3898-3905.
405. Frampton MJE, Law P, Litchfield K, et al. Implications of polygenic risk for personalised colorectal cancer screening. *Ann Oncol*. 2016;27(3):429-434.
406. Pashayan N, Duffy SW, Chowdhury S, et al. Polygenic susceptibility to prostate and breast cancer: Implications for personalised screening. *Br J Cancer*. 2011;104(10):1656-1663.
407. Harrison SM, Dolinsky JS, Knight Johnson AE, et al. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genet Med*. 2017;19(10):1096-1104.
408. Kelly MA, Caleshu C, Morales A, et al. Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. *Genet Med*. 2018;20(3):351-359.
409. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat*. 2018;39(11):1517-1524.
410. ClinGen Sequence Variant Interpretation Working Group. *ClinGen Sequence Variant Interpretation Work Group Recommendations for ACMG/AMP Guideline Criteria Code Modifications Nomenclature*.; 2017.
411. Rehm HL, Berg JS, Brooks LD, et al. ClinGen--the Clinical Genome Resource. *N Engl J Med*. 2015;372(23):2235-2242.
412. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *Am J Hum Genet*. 2018;103(4):474-483.
413. Starita LM, Islam MM, Banerjee T, et al. A Multiplex Homology-Directed DNA Repair Assay Reveals the Impact of More Than 1,000 BRCA1 Missense Substitution Variants on Protein

Function. *Am J Hum Genet*. 2018;103(4):498-508.

414. Andrews KA, Ascher DB, Pires DEV, et al. Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. *J Med Genet*. 2018;55(6):384-394.
415. Precision Prevention and Early Detection Working Group. Precision Prevention and Early Detection Working Group Recommendation - Cancer Prevention and Early Detection in Individuals at High Risk for Cancer. https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/prevention-screening-working-group-report.pdf. Published 2016.
416. Department of Health and Social Care. Chief Medical Officer annual report 2016: Generation Genome. https://www.gov.uk/government/publications/chief-medical-officer-annual-report-2016-generation-genome. Published 2017.
417. Lippi G, Mattiuzzi C, Montagnana M. BRCA population screening for predicting breast cancer: for or against? *Ann Transl Med*. 2017;5(13):275.
418. Lieberman S, Lahad A, Tomer A, Cohen C, Levy-Lahad E, Raz A. Population screening for BRCA1/BRCA2 mutations: lessons from qualitative analysis of the screening experience. *Genet Med*. 2017;19(6):628-634.
419. Long EF, Ganz PA. Cost-effectiveness of Universal BRCA1/2 Screening: Evidence-Based Decision Making. *JAMA Oncol*. 2015;1(9):1217-1218.
420. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*. 2018;359(6378):926-930.
421. McKerrell T, Moreno T, Ponstingl H, et al. Development and validation of a comprehensive genomic diagnostic tool for myeloid malignancies. *Blood*. 2016;128(1):e1-9.
422. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet*. 2014;15(9):585-598.
423. Nik-Zainal S, Alexandrov LB, Wedge DC, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*. 2012;149(5):979-993.
424. Srinivasan M, Sedmak D, Jewell S. Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids. *Am J Pathol*. 2002;161(6):1961-1971.
425. Gilbert MTP, Haselkorn T, Bunce M, et al. The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when? *PLoS One*. 2007;2(6):e537.
426. Genomics England. Sample Handling Guidance. https://www.genomicsengland.co.uk/information-for-gmc-staff/sample-handling-guidance/. Published 2018.

Appendices

Appendix 1 - Tumour categorisation (including for registry and treatment centre-based series) and frequency in MPT series

Table A1 - Tumour categorisation (including for registry and treatment centre-based series) and frequency in MPT series

Tumour category	Occurrences in series	Topographical sites included in category if applicable	Morphological descriptors included in category if applicable
Breast	281		Ductal carcinoma, lobular carcinoma
Colorectal	113	Colon, rectum	
Kidney	83		Clear cell carcinoma, papillary carcinoma, chromophobe carcinoma, oncocytic carcinoma
Non-melanoma skin cancer	67	Any cutaneous site	Basal cell carcinoma, squamous cell carcinoma, Bowens disease
Ovary	58		Carcinoma, endometrioid carcinoma, papillary carcinoma, serous cystadenocarcinoma, mucinous adenocarcinoma, primary serous papillary carcinoma of peritoneum, carcinosarcoma
Endometrium	52		Carcinoma, adenocarcinoma, carcinosarcoma
Melanoma	51	Any cutaneous site	Malignant melanoma, melanoma in situ, superficial spreading melanoma
Thyroid	44		Papillary carcinoma, follicular carcinoma, Hurthle cell carcinoma
Haematological lymphoid	38		Lymphoma, lymphocytic leukaemia, myeloma, hairy cell leukaemia, leukaemia, Waldenstroms macroglobulinaemia
Prostate	22		Adenocarcinoma
Lung	21		Adenocarcinoma, squamous cell carcinoma, small cell.
Paraganglioma	20		
Gastrointestinal stromal tumour	17	Gastric, small bowel, gastrointestinal tract, unspecified.	
Soft tissue Sarcoma	17		Sarcoma, hemangiopericytoma, rhabdomyosarcoma, spindle cell sarcoma, fibrosarcoma, myxoid liposarcoma, liposarcoma, malignant solitary fibrous tumour, leiomyosarcoma, giant cell tumour of tendon sheath, fibromyxosarcoma
Pheochromocytoma	17		
CNS meningioma	14		
Aerodigestive tract	13	Sinus, larynx, nasal cavity, nasopharynx, vocal cord, tongue	Squamous cell carcinoma, small cell carcinoma
Pituitary	13		Pituitary adenoma, prolactinoma, adenoma

Bladder	10		Transitional cell carcinoma, papillary transitional cell carcinoma, papillary urothelial carcinoma, urothelial carcinoma in situ
Pancreatic neuroendocrine tumour	10		Neuroendocrine carcinoma, neuroendocrine tumour, glucagonoma
Central nervous system	10		Glioblastoma, oligodendroglioma, astrocytoma, myxopapillary ependymoma, ganglioglioma
Cervix	8		Squamous cell carcinoma, adenocarcinoma
Gastrointestinal neuroendocrine tumour	8	Appendix, small bowel, large bowel, gastrointestinal tract	Carcinoid (unless non-gastrointestinal site specified)
Testicular	8		Teratoma, seminoma
Central nervous system (nerve sheath)	7		Schwannoma, neurofibroma
Pancreas	7		Solid pseudopapillary tumour, neoplasms
Parathyroid	7		Carcinoma, adenoma
Uveal melanoma	6		
Bone benign	6		Exostoses, osteochondroma, haemangioma
Lung carcinoid	5		
Haematological myeloid	4		Myelogenous leukaemia, myelodysplastic syndromes, myeloproliferative diseases
Ovary sex cord-gonadal stromal	4		Granulosa cell tumour, yolk sac tumour, germ cell tumour, Sertoli leydig tumour
Peripheral nervous system (nerve sheath) benign	4		Neurofibroma, schwannoma
Small bowel	4		
Adrenocortical carcinoma	4		
Central nervous system hemangioblastoma	4		
Kidney oncocytoma	4		
Oesophagus	4		
Colorectal polyps	3	Any lower gastrointestinal site if >10 identified	Serrated adenoma, tubular adenoma, tubulovillous adenoma, hyperplastic polyps, adenomatous polyps, adenoma, adenopapilloma
Salivary gland	3		Acinar cell carcinoma, lymphoepithelial carcinoma
Gastric	3		
Lipoma	3		Angiolipoma
Thymus	3		
Biliary tract	2		Adenocarcinoma, cholangiocarcinoma
Cardiac myxoma	2		Myxoma

Congenital hypertrophy of retinal pigment epithelium	2		
Cutaneous leiomyoma	2		
Desmoid	2		
Fibrofolliculoma	2		
Odontogenic	2		Ameloblastoma, odontogenic tumour
Ovary benign	2		Mucinous cystadenoma, borderline tumours
Pancreas benign	2		
Salivary gland benign	2		
Sebaceous	2		Sebaceous adenoma
Thyroid benign	2		Hurthle cell adenoma, adenoma
Uterine leiomyoma	2		
Uterine sarcoma	2		Sarcoma, leiomyosarcoma
Adrenal adenoma	2		
Bone sarcoma	2		
Colorectal benign	2		
Unknown primary	2		Adenocarcinoma of gastrointestinal origin
Anus	1		Squamous cell carcinoma
Breast phyllodes	1		
Soft tissue benign	1		
Haematological polycythaemia	1		
Haematological thrombocythemia	1		
Kidney angiomyolipoma	1		
Liver benign	1		Adenoma
Lung chondroma	1		
Lung hamartoma	1		
Nerve sheath benign	1		
Ovary neuroendocrine	1		
Penis	1		
Placenta	1		Placental site trophoblastic tumour
Peripheral nervous system (nerve sheath) benign	1		Malignant peripheral nerve sheath tumour
Pulmonary lymphangioleiomyomatosis	1		

Retinoblastoma	1		
Skin benign	1		
Skin Merkel cell	1		Merkell cell tumour
Skin sarcoma	1		Angiosarcoma
Sweat gland	1		Adenocarcinoma
Thyroid medullary	1		Medullary thyroid cancer
Ureter	1		
Vulva	1		
Wilms tumour	1		
Adrenal oncocytoma	1		
Eye Benign	1		Retinal angioma
Eye	0	Retina, conjunctiva, orbit, choroid	Haemangiopericytoma, carcinoma
Lacrimal duct	0		
Liver	0		Hepatocellular carcinoma
Mesothelioma	0	Pleura, peritoneum	
Neuroblastoma	0	Peripheral nerves, ethmoidal sinus, adrenal gland	
Vagina	0		
Vulva	0		

Individual categories only assignable once per individual for purposes of counting tumour frequencies. Neoplasms recorded non-specifically e.g. "cancer" assigned to most likely morphological category for site e.g. breast cancer assigned to Breast not Breast phyllodes. Non-melanoma skin cancer includes basal cell carcinoma and squamous cell carcinoma

Appendix 2 - Comprehensive cancer predisposition gene analysis original and filtered gene list

Table A2 – Comprehensive cancer predisposition gene analysis original and filtered gene list

Gene on original list (n=133)[a]	Gene retained for final list (n=83)[b]
ABCB11	*AIP*
AIP	*ALK*
ALK	*APC*
APC	*ATM*
ATM	*AXIN2*
AXIN2	*BAP1*
BAP1	*BMPR1A*
BLM	*BRCA1*
BMPR1A	*BRCA2*
BRCA1	*BRIP1*
BRCA2	*CDC73*
BRIP1	*CDH1*
BUB1B	*CDK4*
CBL	*CDKN1B*
CDC73	*CDKN2A*
CDH1	*CDKN2B*
CDK4	*CEBPA*
CDKN1B	*CHEK2*
CDKN1C	*CYLD*
CDKN2A	*DDB2*
CDKN2B	*DICER1*
CEBPA	*EGFR*
CEP57	*EPCAM*
CHEK2	*ERCC2*
COL7A1	*ERCC3*
CYLD	*ERCC4*
DDB2	*ERCC5*
DICER1	*EXT1*
DIS3L2	*EXT2*
DKC1	*FH*
DOCK8	*FLCN*
EGFR	*GATA2*
ELANE	*HFE*
EPCAM	*HNF1A*
ERCC2	*KIT*
ERCC3	*MAX*
ERCC4	*MEN1*
ERCC5	*MET*
EXT1	*MLH1*
EXT2	*MSH2*
EZH2	*MSH6*
FAH	*MUTYH*

FANCA	*NF1*
FANCB	*NF2*
FANCC	*NTHL1*
FANCD2	*PALB2*
FANCE	*PDGFRA*
FANCF	*PHOX2B*
FANCG	*PMS2*
FANCI	*POLD1*
FANCL	*POLE*
FANCM	*POLH*
FAS	*PRKAR1A*
FH	*PTCH1*
FLCN	*PTEN*
GATA2	*RAD51C*
GBA	*RAD51D*
GJB2	*RB1*
GPC3	*RET*
HFE	*RHBDF2*
HMBS	*RUNX1*
HNF1A	*SDHA*
HRAS	*SDHAF2*
ITK	*SDHB*
KIT	*SDHC*
MAX	*SDHD*
MEN1	*SERPINA1*
MET	*SMAD4*
MLH1	*SMARCA4*
MSH2	*SMARCB1*
MSH6	*SMARCE1*
MTAP	*SRY*
MUTYH	*STK11*
NBN	*SUFU*
NF1	*TGFBR1*
NF2	*TMEM127*
NSD1	*TP53*
NTHL1	*TSC1*
PALB2	*TSC2*
PDGFRA	*VHL*
PHOX2B	*WT1*
PMS1	*XPA*
PMS2	*XPC*
POLD1	
POLE	
POLH	
PRF1	
PRKAR1A	
PRSS1	
PTCH1	

PTEN
PTPN11
RAD51C
RAD51D
RB1
RECQL4
RET
RHBDF2
RMRP
RUNX1
SBDS
SDHA
SDHAF2
SDHB
SDHC
SDHD
SERPINA1
SH2D1A
SLC25A13
SLX4
SMAD4
SMARCA4
SMARCB1
SMARCE1
SOS1
SRY
STAT3
STK11
SUFU
TERT
TGFBR1
TMEM127
TP53
TRIM37
TSC1
TSC2
UROD
VHL
WAS
WRN
WT1
XPA
XPC

a - Sequenced by Illumina TruSight Cancer panel, appearing in Rahman 20141 or included from further literature review (CDKN2B, NTHL1)
b - Considered consistent with non-syndromic adult presentation where deleterious

Appendix 3 – Gene lists used in analysis for variants in putative novel loci associated with cancer predisposition

Table A3 - Gene list used for analysis of truncating variants based on somatically mutated genes and cancer known CPGs

Gene identifier	Gene name	Canonical transcript	Gene identifier	Gene name	Canonical transcript
ENSG00000115977	*AAK1*	ENST00000409085	ENSG00000184634	*MED12*	ENST00000374080
ENSG00000181409	*AATK*	ENST00000326724	ENSG00000133895	*MEN1*	ENST00000337652
ENSG00000085563	*ABCB1*	ENST00000265724	ENSG00000152595	*MEPE*	ENST00000424957
ENSG00000073734	*ABCB11*	ENST00000263817	ENSG00000105976	*MET*	ENST00000318493
ENSG00000103222	*ABCC1*	ENST00000399410	ENSG00000165819	*METTL3*	ENST00000298717
ENSG00000069431	*ABCC9*	ENST00000261200	ENSG00000168958	*MFF*	ENST00000353339
ENSG00000177465	*ACOT4*	ENST00000326303	ENSG00000204516	*MICB*	ENST00000252229
ENSG00000123983	*ACSL3*	ENST00000357430	ENSG00000155545	*MIER3*	ENST00000381213
ENSG00000075624	*ACTB*	ENST00000331789	ENSG00000076242	*MLH1*	ENST00000231790
ENSG00000184009	*ACTG1*	ENST00000575842	ENSG00000143674	*MLK4*	ENST00000366624
ENSG00000077080	*ACTL6B*	ENST00000160382	ENSG00000171843	*MLLT3*	ENST00000380338
ENSG00000148156	*ACTL7B*	ENST00000374667	ENSG00000169184	*MN1*	ENST00000302326
ENSG00000135503	*ACVR1B*	ENST00000541224	ENSG00000164172	*MOCS2*	ENST00000396954
ENSG00000140955	*ADAD2*	ENST00000268624	ENSG00000005381	*MPO*	ENST00000225275
ENSG00000168594	*ADAM29*	ENST00000359240	ENSG00000150054	*MPP7*	ENST00000337532
ENSG00000138316	*ADAMTS14*	ENST00000373208	ENSG00000132313	*MRPL35*	ENST00000337109
ENSG00000145536	*ADAMTS16*	ENST00000274181	ENSG00000095002	*MSH2*	ENST00000233146
ENSG00000087116	*ADAMTS2*	ENST00000251582	ENSG00000113318	*MSH3*	ENST00000265081
ENSG00000106624	*AEBP1*	ENST00000223357	ENSG00000116062	*MSH6*	ENST00000234420
ENSG00000155966	*AFF2*	ENST00000370460	ENSG00000163132	*MSX1*	ENST00000382723
ENSG00000204149	*AGAP6*	ENST00000412531	ENSG00000099810	*MTAP*	ENST00000380172
ENSG00000144891	*AGTR1*	ENST00000542281	ENSG00000103707	*MTFMT*	ENST00000220058
ENSG00000113492	*AGXT2*	ENST00000231420	ENSG00000198793	*MTOR*	ENST00000361445
ENSG00000110711	*AIP*	ENST00000279146	ENSG00000185499	*MUC1*	ENST00000368395
ENSG00000196581	*AJAP1*	ENST00000378191	ENSG00000169876	*MUC17*	ENST00000306151
ENSG00000129474	*AJUBA*	ENST00000262713	ENSG00000198788	*MUC2*	ENST00000441003
ENSG00000151320	*AKAP6*	ENST00000280979	ENSG00000204544	*MUC21*	ENST00000376296
ENSG00000142208	*AKT1*	ENST00000554581	ENSG00000145113	*MUC4*	ENST00000463781
ENSG00000163631	*ALB*	ENST00000295897	ENSG00000117983	*MUC5B*	ENST00000529681
ENSG00000171094	*ALK*	ENST00000389048	ENSG00000184956	*MUC6*	ENST00000421673
ENSG00000163286	*ALPPL2*	ENST00000295453	ENSG00000171195	*MUC7*	ENST00000413702
ENSG00000139344	*AMDHD1*	ENST00000266736	ENSG00000132781	*MUTYH*	ENST00000372098
ENSG00000184675	*AMER1*	ENST00000330258	ENSG00000110921	*MVK*	ENST00000228510
ENSG00000135409	*AMHR2*	ENST00000257863	ENSG00000118513	*MYB*	ENST00000341911
ENSG00000130812	*ANGPTL6*	ENST00000253109	ENSG00000172936	*MYD88*	ENST00000417037
ENSG00000166839	*ANKDD1A*	ENST00000380230	ENSG00000133020	*MYH8*	ENST00000403437
ENSG00000101745	*ANKRD12*	ENST00000262126	ENSG00000128641	*MYO1B*	ENST00000392318
ENSG00000172014	*ANKRD20A4*	ENST00000357336	ENSG00000173418	*NAA20*	ENST00000334982
ENSG00000148513	*ANKRD30A*	ENST00000361713	ENSG00000186462	*NAP1L2*	ENST00000373517
ENSG00000135976	*ANKRD36*	ENST00000420699	ENSG00000131400	*NAPSA*	ENST00000253719
ENSG00000143401	*ANP32E*	ENST00000314136	ENSG00000144035	*NAT8*	ENST00000272425
ENSG00000134982	*APC*	ENST00000457016	ENSG00000067798	*NAV3*	ENST00000536525
ENSG00000132703	*APCS*	ENST00000255040	ENSG00000104320	*NBN*	ENST00000265433
ENSG00000130203	*APOE*	ENST00000252486	ENSG00000163386	*NBPF10*	ENST00000342960
ENSG00000178878	*APOLD1*	ENST00000326765	ENSG00000243452	*NBPF15*	ENST00000442702

ENSG00000184945	*AQP12A*	ENST00000337801	ENSG00000158092	*NCK1*	ENST00000481752
ENSG00000165269	*AQP7*	ENST00000297988	ENSG00000124151	*NCOA3*	ENST00000371998
ENSG00000103375	*AQP8*	ENST00000219660	ENSG00000141027	*NCOR1*	ENST00000268712
ENSG00000169083	*AR*	ENST00000374690	ENSG00000184983	*NDUFA6*	ENST00000498737
ENSG00000120318	*ARAP3*	ENST00000239440	ENSG00000129559	*NEDD8*	ENST00000250495
ENSG00000163219	*ARHGAP25*	ENST00000409202	ENSG00000100285	*NEFH*	ENST00000310624
ENSG00000117713	*ARID1A*	ENST00000324856	ENSG00000171208	*NETO2*	ENST00000562435
ENSG00000189079	*ARID2*	ENST00000334344	ENSG00000196712	*NF1*	ENST00000358273
ENSG00000228696	*ARL17B*	ENST00000450673	ENSG00000186575	*NF2*	ENST00000338641
ENSG00000163466	*ARPC2*	ENST00000295685	ENSG00000116044	*NFE2L2*	ENST00000397062
ENSG00000140450	*ARRDC4*	ENST00000268042	ENSG00000187566	*NHLRC1*	ENST00000340650
ENSG00000006756	*ARSD*	ENST00000381154	ENSG00000140157	*NIPA2*	ENST00000337451
ENSG00000161664	*ASB16*	ENST00000293414	ENSG00000164190	*NIPBL*	ENST00000282516
ENSG00000164122	*ASB5*	ENST00000296525	ENSG00000167034	*NKX3-1*	ENST00000380871
ENSG00000187855	*ASCL4*	ENST00000342331	ENSG00000087095	*NLK*	ENST00000407008
ENSG00000204653	*ASPDH*	ENST00000389208	ENSG00000158077	*NLRP14*	ENST00000299481
ENSG00000148219	*ASTN2*	ENST00000361209	ENSG00000171487	*NLRP5*	ENST00000390649
ENSG00000143970	*ASXL2*	ENST00000435504	ENSG00000174885	*NLRP6*	ENST00000312165
ENSG00000215915	*ATAD3C*	ENST00000378785	ENSG00000179709	*NLRP8*	ENST00000291971
ENSG00000085978	*ATG16L1*	ENST00000392017	ENSG00000197696	*NMB*	ENST00000394588
ENSG00000149311	*ATM*	ENST00000278616	ENSG00000169251	*NMD3*	ENST00000460469
ENSG00000111676	*ATN1*	ENST00000356654	ENSG00000109255	*NMU*	ENST00000264218
ENSG00000168874	*ATOH8*	ENST00000306279	ENSG00000162408	*NOL9*	ENST00000377705
ENSG00000116039	*ATP6V1B1*	ENST00000234396	ENSG00000146909	*NOM1*	ENST00000275820
ENSG00000166377	*ATP9B*	ENST00000426216	ENSG00000148400	*NOTCH1*	ENST00000277541
ENSG00000085224	*ATRX*	ENST00000373344	ENSG00000134250	*NOTCH2*	ENST00000256646
ENSG00000124788	*ATXN1*	ENST00000244769	ENSG00000213240	*NOTCH2NL*	ENST00000369340
ENSG00000066427	*ATXN3*	ENST00000393287	ENSG00000188747	*NOXA1*	ENST00000341349
ENSG00000105778	*AVL9*	ENST00000318709	ENSG00000056291	*NPFFR2*	ENST00000308744
ENSG00000103126	*AXIN1*	ENST00000262320	ENSG00000135838	*NPL*	ENST00000367553
ENSG00000168646	*AXIN2*	ENST00000307078	ENSG00000181163	*NPM1*	ENST00000296930
ENSG00000166710	*B2M*	ENST00000558401	ENSG00000171246	*NPTX1*	ENST00000306773
ENSG00000198488	*B3GNT6*	ENST00000533140	ENSG00000183971	*NPW*	ENST00000329610
ENSG00000175866	*BAIAP2*	ENST00000321300	ENSG00000181019	*NQO1*	ENST00000320623
ENSG00000163930	*BAP1*	ENST00000460680	ENSG00000213281	*NRAS*	ENST00000369535
ENSG00000127152	*BCL11B*	ENST00000357195	ENSG00000106459	*NRF1*	ENST00000393232
ENSG00000110987	*BCL7A*	ENST00000538010	ENSG00000123572	*NRK*	ENST00000428173
ENSG00000180828	*BHLHE22*	ENST00000321870	ENSG00000165671	*NSD1*	ENST00000439151
ENSG00000197299	*BLM*	ENST00000355112	ENSG00000205309	*NT5M*	ENST00000389022
ENSG00000183682	*BMP8A*	ENST00000331593	ENSG00000065057	*NTHL1*	ENST00000219066
ENSG00000107779	*BMPR1A*	ENST00000372037	ENSG00000074590	*NUAK1*	ENST00000261402
ENSG00000145919	*BOD1*	ENST00000311086	ENSG00000196368	*NUDT11*	ENST00000375992
ENSG00000157764	*BRAF*	ENST00000288602	ENSG00000105245	*NUMBL*	ENST00000252891
ENSG00000012048	*BRCA1*	ENST00000471181	ENSG00000102900	*NUP93*	ENST00000308159
ENSG00000139618	*BRCA2*	ENST00000544455	ENSG00000137804	*NUSAP1*	ENST00000559596
ENSG00000112983	*BRD8*	ENST00000254900	ENSG00000122136	*OBP2A*	ENST00000539850
ENSG00000162670	*BRINP3*	ENST00000367462	ENSG00000154358	*OBSCN*	ENST00000570156
ENSG00000136492	*BRIP1*	ENST00000259008	ENSG00000181781	*ODF3L2*	ENST00000315489
ENSG00000151136	*BTBD11*	ENST00000280758	ENSG00000087263	*OGFOD1*	ENST00000566157
ENSG00000159388	*BTG2*	ENST00000290551	ENSG00000130558	*OLFM1*	ENST00000252854
ENSG00000165810	*BTNL9*	ENST00000327705	ENSG00000116329	*OPRD1*	ENST00000234961
ENSG00000156970	*BUB1B*	ENST00000287598	ENSG00000234560	*OR10G8*	ENST00000431524
ENSG00000005379	*BZRAP1*	ENST00000343736	ENSG00000257019	*OR13C2*	ENST00000542196
ENSG00000171987	*C11orf40*	ENST00000307616	ENSG00000172150	*OR1A2*	ENST00000381951
ENSG00000184601	*C14orf180*	ENST00000557649	ENSG00000197887	*OR1S2*	ENST00000302592
ENSG00000186073	*C15orf41*	ENST00000566621	ENSG00000221938	*OR2A14*	ENST00000408899

ENSG00000187013	*C17orf82*	ENST00000335108	ENSG00000221989	*OR2A2*	ENST00000408979
ENSG00000074842	*C19orf10*	ENST00000262947	ENSG00000188558	*OR2G6*	ENST00000343414
ENSG00000163362	*C1orf106*	ENST00000413687	ENSG00000196071	*OR2L13*	ENST00000366478
ENSG00000173369	*C1QB*	ENST00000314933	ENSG00000196936	*OR2L8*	ENST00000357191
ENSG00000223953	*C1QTNF5*	ENST00000445041	ENSG00000162727	*OR2M5*	ENST00000366476
ENSG00000182326	*C1S*	ENST00000406697	ENSG00000177201	*OR2T12*	ENST00000317996
ENSG00000159239	*C2orf81*	ENST00000290390	ENSG00000196240	*OR2T2*	ENST00000342927
ENSG00000187068	*C3orf70*	ENST00000335012	ENSG00000177212	*OR2T33*	ENST00000318021
ENSG00000174749	*C4orf32*	ENST00000309733	ENSG00000183310	*OR2T34*	ENST00000328782
ENSG00000163633	*C4orf36*	ENST00000473559	ENSG00000196944	*OR2T4*	ENST00000366475
ENSG00000039537	*C6*	ENST00000263413	ENSG00000177462	*OR2T8*	ENST00000319968
ENSG00000112539	*C6orf118*	ENST00000230301	ENSG00000221840	*OR4A5*	ENST00000319760
ENSG00000112936	*C7*	ENST00000313164	ENSG00000181935	*OR4C16*	ENST00000314634
ENSG00000146540	*C7orf50*	ENST00000397098	ENSG00000176547	*OR4C3*	ENST00000319856
ENSG00000157131	*C8A*	ENST00000361249	ENSG00000141194	*OR4D1*	ENST00000268912
ENSG00000213865	*C8orf44*	ENST00000519561	ENSG00000176200	*OR4D11*	ENST00000313253
ENSG00000183784	*C9orf66*	ENST00000382387	ENSG00000182854	*OR4F15*	ENST00000332238
ENSG00000105507	*CABP5*	ENST00000293255	ENSG00000182974	*OR4M2*	ENST00000332663
ENSG00000004948	*CALCR*	ENST00000359558	ENSG00000176294	*OR4N2*	ENST00000315947
ENSG00000108509	*CAMTA2*	ENST00000414043	ENSG00000176895	*OR51A7*	ENST00000359350
ENSG00000064012	*CASP8*	ENST00000358485	ENSG00000184881	*OR51B2*	ENST00000328813
ENSG00000118729	*CASQ2*	ENST00000261448	ENSG00000242180	*OR51B5*	ENST00000300773
ENSG00000067955	*CBFB*	ENST00000412916	ENSG00000176879	*OR51G1*	ENST00000321961
ENSG00000110395	*CBL*	ENST00000264033	ENSG00000176893	*OR51G2*	ENST00000322013
ENSG00000054803	*CBLN4*	ENST00000064571	ENSG00000181609	*OR52D1*	ENST00000322641
ENSG00000122565	*CBX3*	ENST00000337620	ENSG00000176937	*OR52R1*	ENST00000356069
ENSG00000135736	*CCDC102A*	ENST00000258214	ENSG00000172459	*OR5AR1*	ENST00000302969
ENSG00000160994	*CCDC105*	ENST00000292574	ENSG00000198877	*OR5D13*	ENST00000361760
ENSG00000128596	*CCDC136*	ENST00000297788	ENSG00000149133	*OR5F1*	ENST00000278409
ENSG00000248712	*CCDC153*	ENST00000503566	ENSG00000231192	*OR5H1*	ENST00000354565
ENSG00000180376	*CCDC66*	ENST00000394672	ENSG00000236032	*OR5H14*	ENST00000437310
ENSG00000123106	*CCDC91*	ENST00000545336	ENSG00000233412	*OR5H15*	ENST00000356526
ENSG00000142039	*CCDC97*	ENST00000269967	ENSG00000186117	*OR5L1*	ENST00000333973
ENSG00000106178	*CCL24*	ENST00000416943	ENSG00000205030	*OR5L2*	ENST00000378397
ENSG00000110092	*CCND1*	ENST00000227507	ENSG00000174937	*OR5M3*	ENST00000312240
ENSG00000158488	*CD1E*	ENST00000368167	ENSG00000174942	*OR5R1*	ENST00000312253
ENSG00000167775	*CD320*	ENST00000301458	ENSG00000172489	*OR5T3*	ENST00000303059
ENSG00000010610	*CD4*	ENST00000011653	ENSG00000187612	*OR5W2*	ENST00000344514
ENSG00000114013	*CD86*	ENST00000330540	ENSG00000169214	*OR6F1*	ENST00000302084
ENSG00000004897	*CDC27*	ENST00000531206	ENSG00000203757	*OR6K3*	ENST00000368145
ENSG00000134371	*CDC73*	ENST00000367435	ENSG00000198657	*OR8B4*	ENST00000356130
ENSG00000039068	*CDH1*	ENST00000261769	ENSG00000172154	*OR8I2*	ENST00000302124
ENSG00000040731	*CDH10*	ENST00000264463	ENSG00000181689	*OR8K3*	ENST00000312711
ENSG00000154162	*CDH12*	ENST00000382254	ENSG00000181752	*OR8K5*	ENST00000313447
ENSG00000145526	*CDH18*	ENST00000507958	ENSG00000197376	*OR8S1*	ENST00000310194
ENSG00000113100	*CDH9*	ENST00000231021	ENSG00000070882	*OSBPL3*	ENST00000313367
ENSG00000148600	*CDHR1*	ENST00000372117	ENSG00000079156	*OSBPL6*	ENST00000392505
ENSG00000167258	*CDK12*	ENST00000447079	ENSG00000164164	*OTUD4*	ENST00000454497
ENSG00000135446	*CDK4*	ENST00000257904	ENSG00000165588	*OTX2*	ENST00000339475
ENSG00000124762	*CDKN1A*	ENST00000405375	ENSG00000085465	*OVGP1*	ENST00000369732
ENSG00000111276	*CDKN1B*	ENST00000228872	ENSG00000155463	*OXA1L*	ENST00000285848
ENSG00000129757	*CDKN1C*	ENST00000414822	ENSG00000182162	*P2RY8*	ENST00000381297
ENSG00000147889	*CDKN2A*	ENST00000498124	ENSG00000070756	*PABPC1*	ENST00000318607
ENSG00000147883	*CDKN2B*	ENST00000276925	ENSG00000151846	*PABPC3*	ENST00000281589
ENSG00000105352	*CEACAM4*	ENST00000221954	ENSG00000174740	*PABPC5*	ENST00000312600
ENSG00000245848	*CEBPA*	ENST00000498907	ENSG00000083093	*PALB2*	ENST00000261584

ENSG00000093072	*CECR1*	ENST00000399839	ENSG00000149090	*PAMR1*	ENST00000278360
ENSG00000139610	*CELA1*	ENST00000293636	ENSG00000125779	*PANK2*	ENST00000316562
ENSG00000166037	*CEP57*	ENST00000325542	ENSG00000162073	*PAQR4*	ENST00000318782
ENSG00000111642	*CHD4*	ENST00000357008	ENSG00000171053	*PATE1*	ENST00000305738
ENSG00000016391	*CHDH*	ENST00000315251	ENSG00000007372	*PAX6*	ENST00000419022
ENSG00000183765	*CHEK2*	ENST00000382580	ENSG00000163939	*PBRM1*	ENST00000394830
ENSG00000133063	*CHIT1*	ENST00000367229	ENSG00000165494	*PCF11*	ENST00000298281
ENSG00000131873	*CHSY1*	ENST00000254190	ENSG00000056661	*PCGF2*	ENST00000580830
ENSG00000141977	*CIB3*	ENST00000269878	ENSG00000156374	*PCGF6*	ENST00000369847
ENSG00000079432	*CIC*	ENST00000575354	ENSG00000168300	*PCMTD1*	ENST00000360540
ENSG00000113946	*CLDN16*	ENST00000264734	ENSG00000249915	*PDCD6*	ENST00000264933
ENSG00000253958	*CLDN23*	ENST00000519106	ENSG00000134853	*PDGFRA*	ENST00000257290
ENSG00000256660	*CLEC12B*	ENST00000338896	ENSG00000101327	*PDYN*	ENST00000217305
ENSG00000159212	*CLIC6*	ENST00000349499	ENSG00000049246	*PER3*	ENST00000361923
ENSG00000174600	*CMKLR1*	ENST00000312143	ENSG00000154330	*PGM5*	ENST00000396396
ENSG00000176571	*CNBD1*	ENST00000518476	ENSG00000082175	*PGR*	ENST00000325455
ENSG00000155052	*CNTNAP5*	ENST00000431078	ENSG00000164040	*PGRMC2*	ENST00000520121
ENSG00000060718	*COL11A1*	ENST00000370096	ENSG00000156531	*PHF6*	ENST00000332070
ENSG00000164692	*COL1A2*	ENST00000297268	ENSG00000109132	*PHOX2B*	ENST00000226382
ENSG00000169436	*COL22A1*	ENST00000303045	ENSG00000107537	*PHYH*	ENST00000263038
ENSG00000163359	*COL6A3*	ENST00000295550	ENSG00000124102	*PI3*	ENST00000243924
ENSG00000114270	*COL7A1*	ENST00000328333	ENSG00000105229	*PIAS4*	ENST00000262971
ENSG00000173085	*COQ2*	ENST00000311469	ENSG00000121879	*PIK3CA*	ENST00000263967
ENSG00000021826	*CPS1*	ENST00000430249	ENSG00000145675	*PIK3R1*	ENST00000521381
ENSG00000147183	*CPXCR1*	ENST00000276127	ENSG00000170890	*PLA2G1B*	ENST00000308366
ENSG00000203710	*CR1*	ENST00000367049	ENSG00000214456	*PLIN5*	ENST00000381848
ENSG00000134376	*CRB1*	ENST00000367400	ENSG00000106397	*PLOD3*	ENST00000223127
ENSG00000137504	*CREBZF*	ENST00000527447	ENSG00000064933	*PMS1*	ENST00000441310
ENSG00000213145	*CRIP1*	ENST00000330233	ENSG00000122512	*PMS2*	ENST00000265849
ENSG00000257341	*CRIP1*	ENST00000477724	ENSG00000175535	*PNLIP*	ENST00000369221
ENSG00000179979	*CRIPAK*	ENST00000324803	ENSG00000203837	*PNLIPRP3*	ENST00000369230
ENSG00000100122	*CRYBB1*	ENST00000215939	ENSG00000177666	*PNPLA2*	ENST00000336615
ENSG00000168582	*CRYGA*	ENST00000304502	ENSG00000006757	*PNPLA4*	ENST00000381042
ENSG00000169826	*CSGALNACT2*	ENST00000374466	ENSG00000062822	*POLD1*	ENST00000440232
ENSG00000204414	*CSHL1*	ENST00000309894	ENSG00000177084	*POLE*	ENST00000320574
ENSG00000164796	*CSMD3*	ENST00000297405	ENSG00000170734	*POLH*	ENST00000372236
ENSG00000170367	*CST5*	ENST00000304710	ENSG00000221900	*POM121L12*	ENST00000408890
ENSG00000121552	*CSTA*	ENST00000264474	ENSG00000183206	*POTEC*	ENST00000358970
ENSG00000102974	*CTCF*	ENST00000264010	ENSG00000222036	*POTEG*	ENST00000409832
ENSG00000118523	*CTGF*	ENST00000367976	ENSG00000187537	*POTEM*	ENST00000551509
ENSG00000168036	*CTNNB1*	ENST00000349496	ENSG00000170836	*PPM1D*	ENST00000305921
ENSG00000158290	*CUL4B*	ENST00000404115	ENSG00000105568	*PPP2R1A*	ENST00000322088
ENSG00000083799	*CYLD*	ENST00000427738	ENSG00000138814	*PPP3CA*	ENST00000394854
ENSG00000160882	*CYP11B1*	ENST00000292427	ENSG00000119414	*PPP6C*	ENST00000451402
ENSG00000172817	*CYP7B1*	ENST00000310193	ENSG00000116721	*PRAMEF1*	ENST00000332296
ENSG00000152207	*CYSLTR2*	ENST00000282018	ENSG00000251655	*PRB1*	ENST00000500254
ENSG00000126733	*DACH2*	ENST00000373125	ENSG00000121335	*PRB2*	ENST00000389362
ENSG00000100897	*DCAF11*	ENST00000446197	ENSG00000197870	*PRB3*	ENST00000381842
ENSG00000189186	*DCAF8L2*	ENST00000451261	ENSG00000137509	*PRCP*	ENST00000393399
ENSG00000170959	*DCDC1*	ENST00000452803	ENSG00000057657	*PRDM1*	ENST00000369096
ENSG00000133083	*DCLK1*	ENST00000255448	ENSG00000164256	*PRDM9*	ENST00000296682
ENSG00000151065	*DCP1B*	ENST00000280665	ENSG00000085377	*PREP*	ENST00000369110
ENSG00000134574	*DDB2*	ENST00000256996	ENSG00000180644	*PRF1*	ENST00000441259
ENSG00000100523	*DDHD1*	ENST00000323669	ENSG00000146143	*PRIM2*	ENST00000607273
ENSG00000013573	*DDX11*	ENST00000407793	ENSG00000108946	*PRKAR1A*	ENST00000589228
ENSG00000184735	*DDX53*	ENST00000327968	ENSG00000188191	*PRKAR1B*	ENST00000406797

ENSG00000239839	*DEFA3*	ENST00000327857	ENSG00000111218	*PRMT8*	ENST00000382622
ENSG00000176782	*DEFB104A*	ENST00000314265	ENSG00000221961	*PRR21*	ENST00000408934
ENSG00000125788	*DEFB126*	ENST00000382398	ENSG00000116132	*PRRX1*	ENST00000239461
ENSG00000088782	*DEFB127*	ENST00000382388	ENSG00000204983	*PRSS1*	ENST00000311737
ENSG00000124795	*DEK*	ENST00000397239	ENSG00000172382	*PRSS27*	ENST00000302641
ENSG00000100697	*DICER1*	ENST00000526495	ENSG00000105227	*PRX*	ENST00000324001
ENSG00000211448	*DIO2*	ENST00000555750	ENSG00000242221	*PSG2*	ENST00000406487
ENSG00000083520	*DIS3*	ENST00000377767	ENSG00000243137	*PSG4*	ENST00000405312
ENSG00000144535	*DIS3L2*	ENST00000325385	ENSG00000170848	*PSG6*	ENST00000292125
ENSG00000130826	*DKC1*	ENST00000369550	ENSG00000124467	*PSG8*	ENST00000306511
ENSG00000186047	*DLEU7*	ENST00000400393	ENSG00000164985	*PSIP1*	ENST00000380733
ENSG00000161249	*DMKN*	ENST00000339686	ENSG00000108671	*PSMD11*	ENST00000261712
ENSG00000137090	*DMRT1*	ENST00000382276	ENSG00000185920	*PTCH1*	ENST00000331920
ENSG00000163879	*DNALI1*	ENST00000296218	ENSG00000171862	*PTEN*	ENST00000371953
ENSG00000187957	*DNER*	ENST00000341772	ENSG00000165996	*PTPLA*	ENST00000361271
ENSG00000130816	*DNMT1*	ENST00000359526	ENSG00000179295	*PTPN11*	ENST00000351677
ENSG00000119772	*DNMT3A*	ENST00000264709	ENSG00000127947	*PTPN12*	ENST00000248594
ENSG00000134516	*DOCK2*	ENST00000256935	ENSG00000163348	*PYGO2*	ENST00000368457
ENSG00000107099	*DOCK8*	ENST00000453981	ENSG00000163564	*PYHIN1*	ENST00000368140
ENSG00000206052	*DOK6*	ENST00000382713	ENSG00000112531	*QKI*	ENST00000361752
ENSG00000175920	*DOK7*	ENST00000340083	ENSG00000129646	*QRICH2*	ENST00000262765
ENSG00000167130	*DOLPP1*	ENST00000372546	ENSG00000167578	*RAB4B*	ENST00000594800
ENSG00000167261	*DPEP2*	ENST00000412757	ENSG00000136238	*RAC1*	ENST00000356142
ENSG00000121570	*DPPA4*	ENST00000335658	ENSG00000164754	*RAD21*	ENST00000297338
ENSG00000152591	*DSPP*	ENST00000399271	ENSG00000108384	*RAD51C*	ENST00000337432
ENSG00000112679	*DUSP22*	ENST00000344450	ENSG00000185379	*RAD51D*	ENST00000590016
ENSG00000123179	*EBPL*	ENST00000242827	ENSG00000145715	*RASA1*	ENST00000274376
ENSG00000164176	*EDIL3*	ENST00000296591	ENSG00000111344	*RASAL1*	ENST00000546530
ENSG00000136160	*EDNRB*	ENST00000377211	ENSG00000105538	*RASIP1*	ENST00000222145
ENSG00000203666	*EFCAB2*	ENST00000366523	ENSG00000139687	*RB1*	ENST00000267163
ENSG00000146648	*EGFR*	ENST00000275493	ENSG00000182872	*RBM10*	ENST00000377604
ENSG00000120738	*EGR1*	ENST00000239938	ENSG00000244462	*RBM12*	ENST00000374114
ENSG00000173674	*EIF1AX*	ENST00000379607	ENSG00000100461	*RBM23*	ENST00000359890
ENSG00000197561	*ELANE*	ENST00000590230	ENSG00000173933	*RBM4*	ENST00000409406
ENSG00000163435	*ELF3*	ENST00000359651	ENSG00000163694	*RBM47*	ENST00000381793
ENSG00000155849	*ELMO1*	ENST00000310758	ENSG00000147274	*RBMX*	ENST00000320676
ENSG00000162618	*ELTD1*	ENST00000370742	ENSG00000168214	*RBPJ*	ENST00000342295
ENSG00000126749	*EMG1*	ENST00000261406	ENSG00000124232	*RBPJL*	ENST00000343694
ENSG00000143924	*EML4*	ENST00000318522	ENSG00000166965	*RCCD1*	ENST00000394258
ENSG00000163508	*EOMES*	ENST00000295743	ENSG00000160957	*RECQL4*	ENST00000428558
ENSG00000100393	*EP300*	ENST00000263253	ENSG00000115386	*REG1A*	ENST00000233735
ENSG00000183495	*EP400*	ENST00000389561	ENSG00000172023	*REG1B*	ENST00000305089
ENSG00000116016	*EPAS1*	ENST00000263734	ENSG00000172016	*REG3A*	ENST00000393878
ENSG00000129595	*EPB41L4A*	ENST00000261486	ENSG00000143954	*REG3G*	ENST00000272324
ENSG00000119888	*EPCAM*	ENST00000263735	ENSG00000165731	*RET*	ENST00000355710
ENSG00000086289	*EPDR1*	ENST00000199448	ENSG00000223638	*RFPL4A*	ENST00000434937
ENSG00000142627	*EPHA2*	ENST00000358432	ENSG00000132005	*RFX1*	ENST00000254325
ENSG00000080224	*EPHA6*	ENST00000389672	ENSG00000174136	*RGMB*	ENST00000308234
ENSG00000141736	*ERBB2*	ENST00000269571	ENSG00000169629	*RGPD8*	ENST00000302558
ENSG00000065361	*ERBB3*	ENST00000267101	ENSG00000182901	*RGS7*	ENST00000366565
ENSG00000082805	*ERC1*	ENST00000397203	ENSG00000186326	*RGS9BP*	ENST00000334176
ENSG00000104884	*ERCC2*	ENST00000391945	ENSG00000129667	*RHBDF2*	ENST00000313080
ENSG00000163161	*ERCC3*	ENST00000285398	ENSG00000132677	*RHBG*	ENST00000368249
ENSG00000175595	*ERCC4*	ENST00000311895	ENSG00000067560	*RHOA*	ENST00000418115
ENSG00000134899	*ERCC5*	ENST00000355739	ENSG00000143878	*RHOB*	ENST00000272233
ENSG00000187017	*ESPN*	ENST00000377828	ENSG00000119729	*RHOQ*	ENST00000238738

ENSG00000196482	*ESRRG*	ENST00000366937	ENSG00000131941	*RHPN2*	ENST00000254260
ENSG00000182197	*EXT1*	ENST00000378204	ENSG00000187994	*RINL*	ENST00000591812
ENSG00000151348	*EXT2*	ENST00000395673	ENSG00000124784	*RIOK1*	ENST00000379834
ENSG00000188107	*EYS*	ENST00000503581	ENSG00000171136	*RLN3*	ENST00000431365
ENSG00000106462	*EZH2*	ENST00000320356	ENSG00000136104	*RNASEH2B*	ENST00000336617
ENSG00000198734	*F5*	ENST00000367797	ENSG00000181481	*RNF135*	ENST00000328381
ENSG00000103876	*FAH*	ENST00000407106	ENSG00000163162	*RNF149*	ENST00000295317
ENSG00000183688	*FAM101B*	ENST00000329099	ENSG00000189051	*RNF222*	ENST00000399398
ENSG00000182518	*FAM104B*	ENST00000425133	ENSG00000204618	*RNF39*	ENST00000244360
ENSG00000184731	*FAM110C*	ENST00000327669	ENSG00000108375	*RNF43*	ENST00000584437
ENSG00000197798	*FAM118B*	ENST00000533050	ENSG00000114547	*ROPN1B*	ENST00000514116
ENSG00000112584	*FAM120B*	ENST00000476287	ENSG00000156313	*RPGR*	ENST00000378505
ENSG00000156500	*FAM122C*	ENST00000370784	ENSG00000165496	*RPL10L*	ENST00000298283
ENSG00000147724	*FAM135B*	ENST00000395297	ENSG00000116251	*RPL22*	ENST00000234875
ENSG00000182230	*FAM153B*	ENST00000515817	ENSG00000122406	*RPL5*	ENST00000370321
ENSG00000183807	*FAM162B*	ENST00000368557	ENSG00000117676	*RPS6KA1*	ENST00000531382
ENSG00000185442	*FAM174B*	ENST00000327355	ENSG00000144580	*RQCD1*	ENST00000273064
ENSG00000047662	*FAM184B*	ENST00000265018	ENSG00000166592	*RRAD*	ENST00000299759
ENSG00000165837	*FAM194B*	ENST00000298738	ENSG00000124782	*RREB1*	ENST00000379938
ENSG00000122376	*FAM35A*	ENST00000298784	ENSG00000159216	*RUNX1*	ENST00000300305
ENSG00000112773	*FAM46A*	ENST00000320172	ENSG00000079102	*RUNX1T1*	ENST00000436581
ENSG00000183508	*FAM46C*	ENST00000369448	ENSG00000124813	*RUNX2*	ENST00000371438
ENSG00000174016	*FAM46D*	ENST00000538312	ENSG00000186350	*RXRA*	ENST00000481739
ENSG00000170613	*FAM71B*	ENST00000302938	ENSG00000119042	*SATB2*	ENST00000417098
ENSG00000188610	*FAM72B*	ENST00000369390	ENSG00000126524	*SBDS*	ENST00000246868
ENSG00000101447	*FAM83D*	ENST00000217429	ENSG00000185313	*SCN10A*	ENST00000449082
ENSG00000180921	*FAM83H*	ENST00000388913	ENSG00000168356	*SCN11A*	ENST00000302328
ENSG00000186523	*FAM86B1*	ENST00000448228	ENSG00000170616	*SCRT1*	ENST00000332135
ENSG00000145002	*FAM86B2*	ENST00000262365	ENSG00000137575	*SDCBP*	ENST00000260130
ENSG00000183304	*FAM9A*	ENST00000543214	ENSG00000073578	*SDHA*	ENST00000264932
ENSG00000187741	*FANCA*	ENST00000389301	ENSG00000167985	*SDHAF2*	ENST00000301761
ENSG00000181544	*FANCB*	ENST00000398334	ENSG00000117118	*SDHB*	ENST00000375499
ENSG00000158169	*FANCC*	ENST00000289081	ENSG00000143252	*SDHC*	ENST00000367975
ENSG00000144554	*FANCD2*	ENST00000287647	ENSG00000204370	*SDHD*	ENST00000375549
ENSG00000112039	*FANCE*	ENST00000229769	ENSG00000255292	*SDHD*	ENST00000532699
ENSG00000183161	*FANCF*	ENST00000327470	ENSG00000007908	*SELE*	ENST00000333360
ENSG00000221829	*FANCG*	ENST00000378643	ENSG00000075223	*SEMA3C*	ENST00000265361
ENSG00000140525	*FANCI*	ENST00000310775	ENSG00000082684	*SEMA5B*	ENST00000451055
ENSG00000115392	*FANCL*	ENST00000402135	ENSG00000124233	*SEMG1*	ENST00000372781
ENSG00000187790	*FANCM*	ENST00000267430	ENSG00000197249	*SERPINA1*	ENST00000448921
ENSG00000203780	*FANK1*	ENST00000368693	ENSG00000057149	*SERPINB3*	ENST00000283752
ENSG00000026103	*FAS*	ENST00000355740	ENSG00000139718	*SETD1B*	ENST00000267197
ENSG00000083857	*FAT1*	ENST00000441802	ENSG00000181555	*SETD2*	ENST00000409792
ENSG00000086570	*FAT2*	ENST00000261800	ENSG00000174938	*SEZ6L2*	ENST00000308713
ENSG00000165323	*FAT3*	ENST00000298047	ENSG00000168066	*SF1*	ENST00000377387
ENSG00000112787	*FBRSL1*	ENST00000434748	ENSG00000115524	*SF3B1*	ENST00000335508
ENSG00000109670	*FBXW7*	ENST00000281708	ENSG00000118515	*SGK1*	ENST00000367858
ENSG00000203747	*FCGR3A*	ENST00000367969	ENSG00000183918	*SH2D1A*	ENST00000371139
ENSG00000160856	*FCRL3*	ENST00000368184	ENSG00000154447	*SH3RF1*	ENST00000284637
ENSG00000146618	*FERD3L*	ENST00000275461	ENSG00000158352	*SHROOM4*	ENST00000376020
ENSG00000171055	*FEZ2*	ENST00000379245	ENSG00000090402	*SI*	ENST00000264382
ENSG00000066468	*FGFR2*	ENST00000457416	ENSG00000254415	*SIGLEC14*	ENST00000360844
ENSG00000068078	*FGFR3*	ENST00000340107	ENSG00000160584	*SIK3*	ENST00000292055
ENSG00000091483	*FH*	ENST00000366560	ENSG00000112246	*SIM1*	ENST00000369208
ENSG00000134775	*FHOD3*	ENST00000257209	ENSG00000198053	*SIRPA*	ENST00000358771
ENSG00000154803	*FLCN*	ENST00000285071	ENSG00000184302	*SIX6*	ENST00000327720

ENSG00000143631	*FLG*	ENST00000368799	ENSG00000165480	*SKA3*	ENST00000314759
ENSG00000136068	*FLNB*	ENST00000490882	ENSG00000157933	*SKI*	ENST00000378536
ENSG00000122025	*FLT3*	ENST00000241453	ENSG00000081800	*SLC13A1*	ENST00000194130
ENSG00000164694	*FNDC1*	ENST00000297267	ENSG00000004864	*SLC25A13*	ENST00000416240
ENSG00000086205	*FOLH1*	ENST00000256999	ENSG00000091137	*SLC26A4*	ENST00000265715
ENSG00000129514	*FOXA1*	ENST00000250448	ENSG00000014824	*SLC30A9*	ENST00000264451
ENSG00000125798	*FOXA2*	ENST00000419308	ENSG00000139209	*SLC38A4*	ENST00000447411
ENSG00000184492	*FOXD4L1*	ENST00000306507	ENSG00000148482	*SLC39A12*	ENST00000377369
ENSG00000184659	*FOXD4L4*	ENST00000377413	ENSG00000188687	*SLC4A5*	ENST00000377634
ENSG00000178919	*FOXE1*	ENST00000375123	ENSG00000011083	*SLC6A7*	ENST00000230671
ENSG00000164379	*FOXQ1*	ENST00000296839	ENSG00000184564	*SLITRK6*	ENST00000400286
ENSG00000136877	*FPGS*	ENST00000373247	ENSG00000188827	*SLX4*	ENST00000294008
ENSG00000109536	*FRG1*	ENST00000226798	ENSG00000175387	*SMAD2*	ENST00000402690
ENSG00000151474	*FRMD4A*	ENST00000357447	ENSG00000141646	*SMAD4*	ENST00000342988
ENSG00000189139	*FSCB*	ENST00000340446	ENSG00000127616	*SMARCA4*	ENST00000429416
ENSG00000150667	*FSIP1*	ENST00000350221	ENSG00000099956	*SMARCB1*	ENST00000263121
ENSG00000168843	*FSTL5*	ENST00000306100	ENSG00000073584	*SMARCE1*	ENST00000348513
ENSG00000167996	*FTH1*	ENST00000273550	ENSG00000072501	*SMC1A*	ENST00000322213
ENSG00000162613	*FUBP1*	ENST00000370768	ENSG00000108055	*SMC3*	ENST00000361804
ENSG00000157240	*FZD1*	ENST00000287934	ENSG00000116698	*SMG7*	ENST00000507469
ENSG00000104290	*FZD3*	ENST00000240093	ENSG00000188176	*SMTNL2*	ENST00000389313
ENSG00000109158	*GABRA4*	ENST00000264318	ENSG00000132639	*SNAP25*	ENST00000254976
ENSG00000145863	*GABRA6*	ENST00000274545	ENSG00000104976	*SNAPC2*	ENST00000221573
ENSG00000113327	*GABRG2*	ENST00000414552	ENSG00000162804	*SNED1*	ENST00000310397
ENSG00000146276	*GABRR1*	ENST00000454853	ENSG00000100028	*SNRPD3*	ENST00000215829
ENSG00000106105	*GARS*	ENST00000389266	ENSG00000147481	*SNTG1*	ENST00000522124
ENSG00000179348	*GATA2*	ENST00000341105	ENSG00000120669	*SOHLH2*	ENST00000554962
ENSG00000107485	*GATA3*	ENST00000379328	ENSG00000140263	*SORD*	ENST00000267814
ENSG00000177628	*GBA*	ENST00000327247	ENSG00000115904	*SOS1*	ENST00000426016
ENSG00000162654	*GBP4*	ENST00000355754	ENSG00000164736	*SOX17*	ENST00000297316
ENSG00000100116	*GCAT*	ENST00000323205	ENSG00000124766	*SOX4*	ENST00000244745
ENSG00000178795	*GDPD4*	ENST00000315938	ENSG00000125398	*SOX9*	ENST00000245479
ENSG00000103365	*GGA2*	ENST00000309859	ENSG00000105866	*SP4*	ENST00000222584
ENSG00000179168	*GGN*	ENST00000334928	ENSG00000164651	*SP8*	ENST00000418710
ENSG00000123159	*GIPC1*	ENST00000393033	ENSG00000196406	*SPANXD*	ENST00000370515
ENSG00000165474	*GJB2*	ENST00000382844	ENSG00000203923	*SPANXN1*	ENST00000370493
ENSG00000106571	*GLI3*	ENST00000395925	ENSG00000163071	*SPATA18*	ENST00000295213
ENSG00000204007	*GLT6D1*	ENST00000371763	ENSG00000166118	*SPATA19*	ENST00000299140
ENSG00000182327	*GLTPD2*	ENST00000331264	ENSG00000141255	*SPATA22*	ENST00000573128
ENSG00000105373	*GLTSCR2*	ENST00000246802	ENSG00000174015	*SPERT*	ENST00000310521
ENSG00000104499	*GML*	ENST00000220940	ENSG00000133104	*SPG20*	ENST00000451493
ENSG00000088256	*GNA11*	ENST00000078429	ENSG00000153820	*SPHKAP*	ENST00000392056
ENSG00000156052	*GNAQ*	ENST00000286548	ENSG00000147059	*SPIN2A*	ENST00000374908
ENSG00000087460	*GNAS*	ENST00000371100	ENSG00000121067	*SPOP*	ENST00000393331
ENSG00000172380	*GNG12*	ENST00000370982	ENSG00000169474	*SPRR1A*	ENST00000307122
ENSG00000215405	*GOLGA6L6*	ENST00000427390	ENSG00000187678	*SPRY4*	ENST00000344120
ENSG00000147257	*GPC3*	ENST00000394299	ENSG00000163554	*SPTA1*	ENST00000368147
ENSG00000183098	*GPC6*	ENST00000377047	ENSG00000167978	*SRRM2*	ENST00000301740
ENSG00000146360	*GPR6*	ENST00000275169	ENSG00000184895	*SRY*	ENST00000383070
ENSG00000204175	*GPRIN2*	ENST00000374314	ENSG00000157216	*SSBP3*	ENST00000371320
ENSG00000132522	*GPS2*	ENST00000380728	ENSG00000101972	*STAG2*	ENST00000218089
ENSG00000213654	*GPSM3*	ENST00000375040	ENSG00000168610	*STAT3*	ENST00000264657
ENSG00000109519	*GRPEL1*	ENST00000264954	ENSG00000118046	*STK11*	ENST00000326873
ENSG00000215203	*GRXCR1*	ENST00000399770	ENSG00000204344	*STK19*	ENST00000375333
ENSG00000244067	*GSTA2*	ENST00000493422	ENSG00000107882	*SUFU*	ENST00000369902
ENSG00000100577	*GSTZ1*	ENST00000216465	ENSG00000173597	*SULT1B1*	ENST00000310613

ENSG00000148702	*HABP2*	ENST00000351270	ENSG00000178691	*SUZ12*	ENST00000322652
ENSG00000131373	*HACL1*	ENST00000321169	ENSG00000122012	*SV2C*	ENST00000502798
ENSG00000223609	*HBD*	ENST00000380299	ENSG00000131018	*SYNE1*	ENST00000367255
ENSG00000196565	*HBG2*	ENST00000380259	ENSG00000176438	*SYNE3*	ENST00000334258
ENSG00000155393	*HEATR3*	ENST00000299192	ENSG00000149043	*SYT8*	ENST00000381968
ENSG00000165338	*HECTD2*	ENST00000298068	ENSG00000148835	*TAF5*	ENST00000369839
ENSG00000002746	*HECW1*	ENST00000395891	ENSG00000169777	*TAS2R1*	ENST00000382492
ENSG00000010704	*HFE*	ENST00000357618	ENSG00000121318	*TAS2R10*	ENST00000240619
ENSG00000182218	*HHIPL1*	ENST00000330710	ENSG00000127362	*TAS2R3*	ENST00000247879
ENSG00000114455	*HHLA2*	ENST00000357759	ENSG00000255374	*TAS2R43*	ENST00000531678
ENSG00000148110	*HIATL1*	ENST00000375344	ENSG00000122145	*TBX22*	ENST00000373294
ENSG00000168298	*HIST1H1E*	ENST00000304218	ENSG00000135111	*TBX3*	ENST00000257566
ENSG00000164508	*HIST1H2AA*	ENST00000297012	ENSG00000121075	*TBX4*	ENST00000240335
ENSG00000185130	*HIST1H2BL*	ENST00000377401	ENSG00000204065	*TCEAL5*	ENST00000372680
ENSG00000124693	*HIST1H3B*	ENST00000244661	ENSG00000154582	*TCEB1*	ENST00000518127
ENSG00000206503	*HLA-A*	ENST00000396634	ENSG00000113649	*TCERG1*	ENST00000296702
ENSG00000234745	*HLA-B*	ENST00000412585	ENSG00000140262	*TCF12*	ENST00000438423
ENSG00000198502	*HLA-DRB5*	ENST00000374975	ENSG00000148737	*TCF7L2*	ENST00000543371
ENSG00000256269	*HMBS*	ENST00000278715	ENSG00000163060	*TEKT4*	ENST00000295201
ENSG00000205581	*HMGN1*	ENST00000380749	ENSG00000164362	*TERT*	ENST00000310581
ENSG00000135100	*HNF1A*	ENST00000257555	ENSG00000168769	*TET2*	ENST00000540549
ENSG00000179172	*HNRNPCL1*	ENST00000317869	ENSG00000106799	*TGFBR1*	ENST00000374994
ENSG00000165119	*HNRNPK*	ENST00000376263	ENSG00000163513	*TGFBR2*	ENST00000359013
ENSG00000099783	*HNRNPM*	ENST00000325495	ENSG00000153779	*TGIF2LX*	ENST00000561129
ENSG00000163755	*HPS3*	ENST00000296051	ENSG00000173451	*THAP2*	ENST00000308086
ENSG00000174775	*HRAS*	ENST00000451590	ENSG00000159445	*THEM4*	ENST00000368814
ENSG00000196196	*HRCT1*	ENST00000354323	ENSG00000196407	*THEM5*	ENST00000368817
ENSG00000108786	*HSD17B1*	ENST00000585807	ENSG00000144229	*THSD7B*	ENST00000272643
ENSG00000102878	*HSF4*	ENST00000264009	ENSG00000038295	*TLL1*	ENST00000061240
ENSG00000115541	*HSPE1*	ENST00000233893	ENSG00000136869	*TLR4*	ENST00000355622
ENSG00000138413	*IDH1*	ENST00000415913	ENSG00000135956	*TMEM127*	ENST00000258439
ENSG00000182054	*IDH2*	ENST00000330062	ENSG00000206432	*TMEM200C*	ENST00000581347
ENSG00000127415	*IDUA*	ENST00000247933	ENSG00000119777	*TMEM214*	ENST00000238788
ENSG00000162783	*IER5*	ENST00000367577	ENSG00000234224	*TMEM229A*	ENST00000455783
ENSG00000142089	*IFITM3*	ENST00000399808	ENSG00000155099	*TMEM55A*	ENST00000285419
ENSG00000254709	*IGLL5*	ENST00000526893	ENSG00000137747	*TMPRSS13*	ENST00000524993
ENSG00000136634	*IL10*	ENST00000423557	ENSG00000205542	*TMSB4X*	ENST00000380636
ENSG00000115607	*IL18RAP*	ENST00000264260	ENSG00000133687	*TMTC1*	ENST00000539277
ENSG00000016402	*IL20RA*	ENST00000316649	ENSG00000123610	*TNFAIP6*	ENST00000243347
ENSG00000008517	*IL32*	ENST00000525643	ENSG00000157873	*TNFRSF14*	ENST00000355716
ENSG00000134352	*IL6ST*	ENST00000381298	ENSG00000243509	*TNFRSF6B*	ENST00000369996
ENSG00000153487	*ING1*	ENST00000375774	ENSG00000106952	*TNFSF8*	ENST00000223795
ENSG00000068745	*IP6K2*	ENST00000328631	ENSG00000168884	*TNIP2*	ENST00000315423
ENSG00000168310	*IRF2*	ENST00000393593	ENSG00000186283	*TOR3A*	ENST00000367627
ENSG00000137265	*IRF4*	ENST00000380956	ENSG00000141510	*TP53*	ENST00000269305
ENSG00000133124	*IRS4*	ENST00000372129	ENSG00000124251	*TP53TG5*	ENST00000372726
ENSG00000177508	*IRX3*	ENST00000329734	ENSG00000115705	*TPO*	ENST00000345913
ENSG00000113263	*ITK*	ENST00000422843	ENSG00000171368	*TPPP*	ENST00000360578
ENSG00000009765	*IYD*	ENST00000229447	ENSG00000178928	*TPRX1*	ENST00000322175
ENSG00000081692	*JMJD4*	ENST00000366758	ENSG00000095917	*TPSD1*	ENST00000211076
ENSG00000152409	*JMY*	ENST00000396137	ENSG00000169902	*TPST1*	ENST00000304842
ENSG00000186994	*KANK3*	ENST00000330915	ENSG00000166157	*TPTE*	ENST00000361285
ENSG00000083168	*KAT6A*	ENST00000396930	ENSG00000132958	*TPTE2*	ENST00000400230
ENSG00000234438	*KBTBD13*	ENST00000432196	ENSG00000131323	*TRAF3*	ENST00000560371
ENSG00000180509	*KCNE1*	ENST00000337385	ENSG00000174599	*TRAM1L1*	ENST00000310754
ENSG00000124780	*KCNK17*	ENST00000373231	ENSG00000112195	*TREML2*	ENST00000483722

ENSG00000164626	*KCNK5*	ENST00000359534	ENSG00000072657	*TRHDE*	ENST00000261180
ENSG00000143603	*KCNN3*	ENST00000271915	ENSG00000108395	*TRIM37*	ENST00000262294
ENSG00000162687	*KCNT2*	ENST00000294725	ENSG00000150244	*TRIM48*	ENST00000417545
ENSG00000215262	*KCNU1*	ENST00000399881	ENSG00000147573	*TRIM55*	ENST00000315962
ENSG00000155729	*KCTD18*	ENST00000359878	ENSG00000162722	*TRIM58*	ENST00000366481
ENSG00000136636	*KCTD3*	ENST00000259154	ENSG00000179046	*TRIML2*	ENST00000512729
ENSG00000126012	*KDM5C*	ENST00000375401	ENSG00000100106	*TRIOBP*	ENST00000406386
ENSG00000147050	*KDM6A*	ENST00000377967	ENSG00000165699	*TSC1*	ENST00000298552
ENSG00000079999	*KEAP1*	ENST00000171111	ENSG00000103197	*TSC2*	ENST00000219476
ENSG00000197993	*KEL*	ENST00000355265	ENSG00000196428	*TSC22D2*	ENST00000361875
ENSG00000235750	*KIAA0040*	ENST00000545251	ENSG00000126467	*TSKS*	ENST00000246801
ENSG00000134313	*KIDINS220*	ENST00000256707	ENSG00000231738	*TSPAN19*	ENST00000532498
ENSG00000157404	*KIT*	ENST00000288135	ENSG00000155657	*TTN*	ENST00000589042
ENSG00000109787	*KLF3*	ENST00000261438	ENSG00000104723	*TUSC3*	ENST00000503731
ENSG00000102554	*KLF5*	ENST00000377687	ENSG00000077498	*TYR*	ENST00000263321
ENSG00000205810	*KLRC3*	ENST00000381903	ENSG00000092445	*TYRO3*	ENST00000263798
ENSG00000055609	*KMT2C*	ENST00000262189	ENSG00000160201	*U2AF1*	ENST00000291552
ENSG00000167548	*KMT2D*	ENST00000301067	ENSG00000130560	*UBAC1*	ENST00000371756
ENSG00000171798	*KNDC1*	ENST00000304613	ENSG00000077721	*UBE2A*	ENST00000371558
ENSG00000133703	*KRAS*	ENST00000256078	ENSG00000158062	*UBXN11*	ENST00000374222
ENSG00000171346	*KRT15*	ENST00000254043	ENSG00000135220	*UGT2A3*	ENST00000251566
ENSG00000172867	*KRT2*	ENST00000309680	ENSG00000083290	*ULK2*	ENST00000395544
ENSG00000139648	*KRT71*	ENST00000267119	ENSG00000168038	*ULK4*	ENST00000301831
ENSG00000161849	*KRT84*	ENST00000257951	ENSG00000115446	*UNC50*	ENST00000357765
ENSG00000221859	*KRTAP10-10*	ENST00000380095	ENSG00000169021	*UQCRFS1*	ENST00000304863
ENSG00000243489	*KRTAP10-11*	ENST00000334670	ENSG00000126088	*UROD*	ENST00000246337
ENSG00000205441	*KRTAP10-7*	ENST00000380102	ENSG00000143258	*USP21*	ENST00000368002
ENSG00000188581	*KRTAP1-1*	ENST00000306271	ENSG00000131864	*USP29*	ENST00000254181
ENSG00000187026	*KRTAP21-2*	ENST00000333892	ENSG00000106346	*USP42*	ENST00000306177
ENSG00000214518	*KRTAP2-2*	ENST00000398477	ENSG00000181408	*UTS2R*	ENST00000313135
ENSG00000188694	*KRTAP24-1*	ENST00000340345	ENSG00000177504	*VCX2*	ENST00000317103
ENSG00000206107	*KRTAP27-1*	ENST00000382835	ENSG00000150630	*VEGFC*	ENST00000280193
ENSG00000212721	*KRTAP4-11*	ENST00000391413	ENSG00000134086	*VHL*	ENST00000256474
ENSG00000198271	*KRTAP4-5*	ENST00000343246	ENSG00000178201	*VN1R1*	ENST00000321039
ENSG00000240871	*KRTAP4-7*	ENST00000391417	ENSG00000188730	*VWC2*	ENST00000340652
ENSG00000205869	*KRTAP5-1*	ENST00000382171	ENSG00000015285	*WAS*	ENST00000376701
ENSG00000185940	*KRTAP5-5*	ENST00000399676	ENSG00000239779	*WBP1*	ENST00000233615
ENSG00000244411	*KRTAP5-7*	ENST00000398536	ENSG00000119333	*WDR34*	ENST00000372715
ENSG00000239886	*KRTAP9-2*	ENST00000377721	ENSG00000174776	*WDR49*	ENST00000308378
ENSG00000241595	*KRTAP9-4*	ENST00000334109	ENSG00000206530	*WDR52*	ENST00000393845
ENSG00000103642	*LACTB*	ENST00000261893	ENSG00000060237	*WNK1*	ENST00000315939
ENSG00000107929	*LARP4B*	ENST00000316157	ENSG00000002745	*WNT16*	ENST00000222462
ENSG00000196734	*LCE1B*	ENST00000360090	ENSG00000105989	*WNT2*	ENST00000265441
ENSG00000240386	*LCE1F*	ENST00000334371	ENSG00000165392	*WRN*	ENST00000298139
ENSG00000187173	*LCE2A*	ENST00000368779	ENSG00000184937	*WT1*	ENST00000332351
ENSG00000163202	*LCE3D*	ENST00000368787	ENSG00000136936	*XPA*	ENST00000375128
ENSG00000169744	*LDB2*	ENST00000304523	ENSG00000154767	*XPC*	ENST00000285021
ENSG00000182909	*LENG9*	ENST00000333834	ENSG00000015532	*XYLT2*	ENST00000017003
ENSG00000168924	*LETM1*	ENST00000302787	ENSG00000174851	*YIF1A*	ENST00000376901
ENSG00000050426	*LETMD1*	ENST00000418425	ENSG00000182223	*ZAR1*	ENST00000327939
ENSG00000138039	*LHCGR*	ENST00000294954	ENSG00000169064	*ZBBX*	ENST00000455345
ENSG00000182508	*LHFPL1*	ENST00000371968	ENSG00000181722	*ZBTB20*	ENST00000474710
ENSG00000239998	*LILRA2*	ENST00000251377	ENSG00000160685	*ZBTB7B*	ENST00000417934
ENSG00000182541	*LIMK2*	ENST00000340552	ENSG00000178199	*ZC3H12D*	ENST00000409806
ENSG00000101670	*LIPG*	ENST00000261292	ENSG00000177764	*ZCCHC3*	ENST00000382352
ENSG00000074695	*LMAN1*	ENST00000251047	ENSG00000156599	*ZDHHC5*	ENST00000287169

ENSG00000170807	*LMOD2*	ENST00000458573	ENSG00000169554	*ZEB2*	ENST00000558170
ENSG00000164715	*LMTK2*	ENST00000297293	ENSG00000140836	*ZFHX3*	ENST00000268489
ENSG00000203782	*LOR*	ENST00000368742	ENSG00000185650	*ZFP36L1*	ENST00000439696
ENSG00000117600	*LPPR4*	ENST00000370185	ENSG00000152518	*ZFP36L2*	ENST00000282388
ENSG00000144749	*LRIG1*	ENST00000273261	ENSG00000179588	*ZFPM1*	ENST00000319555
ENSG00000120256	*LRP11*	ENST00000239367	ENSG00000122515	*ZMIZ2*	ENST00000309315
ENSG00000158113	*LRRC43*	ENST00000339777	ENSG00000160321	*ZNF208*	ENST00000397126
ENSG00000131409	*LRRC4B*	ENST00000599957	ENSG00000267508	*ZNF285*	ENST00000330997
ENSG00000148948	*LRRC4C*	ENST00000278198	ENSG00000105136	*ZNF419*	ENST00000424930
ENSG00000171017	*LRRC8E*	ENST00000306708	ENSG00000229676	*ZNF492*	ENST00000456783
ENSG00000162620	*LRRIQ3*	ENST00000354431	ENSG00000197363	*ZNF517*	ENST00000359971
ENSG00000125872	*LRRN4*	ENST00000378858	ENSG00000197701	*ZNF595*	ENST00000526473
ENSG00000168056	*LTBP3*	ENST00000301873	ENSG00000167962	*ZNF598*	ENST00000431526
ENSG00000062524	*LTK*	ENST00000263800	ENSG00000257591	*ZNF625*	ENST00000439556
ENSG00000139329	*LUM*	ENST00000266718	ENSG00000188171	*ZNF626*	ENST00000601440
ENSG00000187398	*LUZP2*	ENST00000336930	ENSG00000197483	*ZNF628*	ENST00000598519
ENSG00000099949	*LZTR1*	ENST00000215739	ENSG00000196109	*ZNF676*	ENST00000397121
ENSG00000061337	*LZTS1*	ENST00000381569	ENSG00000197123	*ZNF679*	ENST00000421025
ENSG00000099866	*MADCAM1*	ENST00000215637	ENSG00000196946	*ZNF705A*	ENST00000359286
ENSG00000110514	*MADD*	ENST00000311027	ENSG00000182141	*ZNF708*	ENST00000356929
ENSG00000177689	*MAGEB10*	ENST00000356790	ENSG00000141579	*ZNF750*	ENST00000269394
ENSG00000099399	*MAGEB2*	ENST00000378988	ENSG00000198146	*ZNF770*	ENST00000356321
ENSG00000155495	*MAGEC1*	ENST00000285879	ENSG00000196456	*ZNF775*	ENST00000329630
ENSG00000147676	*MAL2*	ENST00000276681	ENSG00000170396	*ZNF804A*	ENST00000302277
ENSG00000130479	*MAP1S*	ENST00000324096	ENSG00000182348	*ZNF804B*	ENST00000333190
ENSG00000065559	*MAP2K4*	ENST00000353533	ENSG00000204514	*ZNF814*	ENST00000435989
ENSG00000076984	*MAP2K7*	ENST00000397979	ENSG00000257446	*ZNF878*	ENST00000547628
ENSG00000095015	*MAP3K1*	ENST00000399503	ENSG00000234284	*ZNF879*	ENST00000444149
ENSG00000135525	*MAP7*	ENST00000454590	ENSG00000221923	*ZNF880*	ENST00000422689
ENSG00000100030	*MAPK1*	ENST00000215832	ENSG00000213973	*ZNF99*	ENST00000596209
ENSG00000186868	*MAPT*	ENST00000344290	ENSG00000188372	*ZP3*	ENST00000394857
ENSG00000007047	*MARK4*	ENST00000262891	ENSG00000042813	*ZPBP*	ENST00000046087
ENSG00000125952	*MAX*	ENST00000358664	ENSG00000131848	*ZSCAN5A*	ENST00000587340
ENSG00000166987	*MBD6*	ENST00000355673	ENSG00000122952	*ZWINT*	ENST00000373944
ENSG00000213920	*MDP1*	ENST00000288087			

Table A4 - Gene list used for analysis of truncating variants based on somatically mutated genes and cancer known CPGs – Refined with LOFTOOL

Gene identifier	Gene name	Canonical transcript	Gene identifier	Gene name	Canonical transcript
ENSG00000115977	*AAK1*	ENST00000409085	ENSG00000061337	*LZTS1*	ENST00000381569
ENSG00000181409	*AATK*	ENST00000326724	ENSG00000099866	*MADCAM1*	ENST00000215637
ENSG00000085563	*ABCB1*	ENST00000265724	ENSG00000177689	*MAGEB10*	ENST00000356790
ENSG00000073734	*ABCB11*	ENST00000263817	ENSG00000147676	*MAL2*	ENST00000276681
ENSG00000103222	*ABCC1*	ENST00000399410	ENSG00000130479	*MAP1S*	ENST00000324096
ENSG00000069431	*ABCC9*	ENST00000261200	ENSG00000065559	*MAP2K4*	ENST00000353533
ENSG00000075624	*ACTB*	ENST00000331789	ENSG00000076984	*MAP2K7*	ENST00000397979
ENSG00000184009	*ACTG1*	ENST00000575842	ENSG00000100030	*MAPK1*	ENST00000215832
ENSG00000148156	*ACTL7B*	ENST00000374667	ENSG00000186868	*MAPT*	ENST00000344290
ENSG00000135503	*ACVR1B*	ENST00000541224	ENSG00000007047	*MARK4*	ENST00000262891
ENSG00000140955	*ADAD2*	ENST00000268624	ENSG00000125952	*MAX*	ENST00000358664
ENSG00000087116	*ADAMTS2*	ENST00000251582	ENSG00000166987	*MBD6*	ENST00000355673
ENSG00000106624	*AEBP1*	ENST00000223357	ENSG00000184634	*MED12*	ENST00000374080
ENSG00000155966	*AFF2*	ENST00000370460	ENSG00000133895	*MEN1*	ENST00000337652
ENSG00000204149	*AGAP6*	ENST00000412531	ENSG00000076242	*MLH1*	ENST00000231790
ENSG00000110711	*AIP*	ENST00000279146	ENSG00000143674	*MLK4*	ENST00000366624
ENSG00000196581	*AJAP1*	ENST00000378191	ENSG00000171843	*MLLT3*	ENST00000380338
ENSG00000129474	*AJUBA*	ENST00000262713	ENSG00000169184	*MN1*	ENST00000302326
ENSG00000171094	*ALK*	ENST00000389048	ENSG00000095002	*MSH2*	ENST00000233146
ENSG00000163286	*ALPPL2*	ENST00000295453	ENSG00000113318	*MSH3*	ENST00000265081
ENSG00000184675	*AMER1*	ENST00000330258	ENSG00000116062	*MSH6*	ENST00000234420
ENSG00000135409	*AMHR2*	ENST00000257863	ENSG00000163132	*MSX1*	ENST00000382723
ENSG00000130812	*ANGPTL6*	ENST00000253109	ENSG00000198793	*MTOR*	ENST00000361445
ENSG00000172014	*ANKRD20A4*	ENST00000357336	ENSG00000198788	*MUC2*	ENST00000441003
ENSG00000135976	*ANKRD36*	ENST00000420699	ENSG00000117983	*MUC5B*	ENST00000529681
ENSG00000134982	*APC*	ENST00000457016	ENSG00000132781	*MUTYH*	ENST00000372098
ENSG00000130203	*APOE*	ENST00000252486	ENSG00000110921	*MVK*	ENST00000228510
ENSG00000178878	*APOLD1*	ENST00000326765	ENSG00000133020	*MYH8*	ENST00000403437
ENSG00000184945	*AQP12A*	ENST00000337801	ENSG00000128641	*MYO1B*	ENST00000392318
ENSG00000165269	*AQP7*	ENST00000297988	ENSG00000067798	*NAV3*	ENST00000536525
ENSG00000169083	*AR*	ENST00000374690	ENSG00000163386	*NBPF10*	ENST00000342960
ENSG00000117713	*ARID1A*	ENST00000324856	ENSG00000129559	*NEDD8*	ENST00000250495
ENSG00000163466	*ARPC2*	ENST00000295685	ENSG00000100285	*NEFH*	ENST00000310624
ENSG00000006756	*ARSD*	ENST00000381154	ENSG00000196712	*NF1*	ENST00000358273
ENSG00000187855	*ASCL4*	ENST00000342331	ENSG00000186575	*NF2*	ENST00000338641
ENSG00000143970	*ASXL2*	ENST00000435504	ENSG00000116044	*NFE2L2*	ENST00000397062
ENSG00000168874	*ATOH8*	ENST00000306279	ENSG00000187566	*NHLRC1*	ENST00000340650
ENSG00000166377	*ATP9B*	ENST00000426216	ENSG00000140157	*NIPA2*	ENST00000337451
ENSG00000085224	*ATRX*	ENST00000373344	ENSG00000164190	*NIPBL*	ENST00000282516
ENSG00000124788	*ATXN1*	ENST00000244769	ENSG00000167034	*NKX3-1*	ENST00000380871
ENSG00000105778	*AVL9*	ENST00000318709	ENSG00000087095	*NLK*	ENST00000407008
ENSG00000103126	*AXIN1*	ENST00000262320	ENSG00000148400	*NOTCH1*	ENST00000277541
ENSG00000168646	*AXIN2*	ENST00000307078	ENSG00000134250	*NOTCH2*	ENST00000256646
ENSG00000166710	*B2M*	ENST00000558401	ENSG00000171246	*NPTX1*	ENST00000306773
ENSG00000198488	*B3GNT6*	ENST00000533140	ENSG00000183971	*NPW*	ENST00000329610
ENSG00000127152	*BCL11B*	ENST00000357195	ENSG00000213281	*NRAS*	ENST00000369535
ENSG00000180828	*BHLHE22*	ENST00000321870	ENSG00000106459	*NRF1*	ENST00000393232
ENSG00000183682	*BMP8A*	ENST00000331593	ENSG00000123572	*NRK*	ENST00000428173
ENSG00000107779	*BMPR1A*	ENST00000372037	ENSG00000165671	*NSD1*	ENST00000439151
ENSG00000157764	*BRAF*	ENST00000288602	ENSG00000196368	*NUDT11*	ENST00000375992
ENSG00000012048	*BRCA1*	ENST00000471181	ENSG00000105245	*NUMBL*	ENST00000252891

ENSG00000139618	*BRCA2*	ENST00000544455	ENSG00000154358	*OBSCN*	ENST00000570156
ENSG00000162670	*BRINP3*	ENST00000367462	ENSG00000181781	*ODF3L2*	ENST00000315489
ENSG00000163362	*C1orf106*	ENST00000413687	ENSG00000130558	*OLFM1*	ENST00000252854
ENSG00000173369	*C1QB*	ENST00000314933	ENSG00000196071	*OR2L13*	ENST00000366478
ENSG00000182326	*C1S*	ENST00000406697	ENSG00000177462	*OR2T8*	ENST00000319968
ENSG00000159239	*C2orf81*	ENST00000290390	ENSG00000242180	*OR51B5*	ENST00000300773
ENSG00000174749	*C4orf32*	ENST00000309733	ENSG00000165588	*OTX2*	ENST00000339475
ENSG00000183784	*C9orf66*	ENST00000382387	ENSG00000182162	*P2RY8*	ENST00000381297
ENSG00000067955	*CBFB*	ENST00000412916	ENSG00000070756	*PABPC1*	ENST00000318607
ENSG00000054803	*CBLN4*	ENST00000064571	ENSG00000174740	*PABPC5*	ENST00000312600
ENSG00000135736	*CCDC102A*	ENST00000258214	ENSG00000149090	*PAMR1*	ENST00000278360
ENSG00000128596	*CCDC136*	ENST00000297788	ENSG00000125779	*PANK2*	ENST00000316562
ENSG00000110092	*CCND1*	ENST00000227507	ENSG00000007372	*PAX6*	ENST00000419022
ENSG00000167775	*CD320*	ENST00000301458	ENSG00000163939	*PBRM1*	ENST00000394830
ENSG00000114013	*CD86*	ENST00000330540	ENSG00000165494	*PCF11*	ENST00000298281
ENSG00000134371	*CDC73*	ENST00000367435	ENSG00000134853	*PDGFRA*	ENST00000257290
ENSG00000167258	*CDK12*	ENST00000447079	ENSG00000101327	*PDYN*	ENST00000217305
ENSG00000124762	*CDKN1A*	ENST00000405375	ENSG00000082175	*PGR*	ENST00000325455
ENSG00000111276	*CDKN1B*	ENST00000228872	ENSG00000164040	*PGRMC2*	ENST00000520121
ENSG00000129757	*CDKN1C*	ENST00000414822	ENSG00000156531	*PHF6*	ENST00000332070
ENSG00000147889	*CDKN2A*	ENST00000498124	ENSG00000109132	*PHOX2B*	ENST00000226382
ENSG00000093072	*CECR1*	ENST00000399839	ENSG00000107537	*PHYH*	ENST00000263038
ENSG00000111642	*CHD4*	ENST00000357008	ENSG00000105229	*PIAS4*	ENST00000262971
ENSG00000131873	*CHSY1*	ENST00000254190	ENSG00000214456	*PLIN5*	ENST00000381848
ENSG00000079432	*CIC*	ENST00000575354	ENSG00000106397	*PLOD3*	ENST00000223127
ENSG00000113946	*CLDN16*	ENST00000264734	ENSG00000064933	*PMS1*	ENST00000441310
ENSG00000159212	*CLIC6*	ENST00000349499	ENSG00000183206	*POTEC*	ENST00000358970
ENSG00000060718	*COL11A1*	ENST00000370096	ENSG00000222036	*POTEG*	ENST00000409832
ENSG00000164692	*COL1A2*	ENST00000297268	ENSG00000138814	*PPP3CA*	ENST00000394854
ENSG00000169436	*COL22A1*	ENST00000303045	ENSG00000197870	*PRB3*	ENST00000381842
ENSG00000163359	*COL6A3*	ENST00000295550	ENSG00000057657	*PRDM1*	ENST00000369096
ENSG00000114270	*COL7A1*	ENST00000328333	ENSG00000180644	*PRF1*	ENST00000441259
ENSG00000021826	*CPS1*	ENST00000430249	ENSG00000146143	*PRIM2*	ENST00000607273
ENSG00000203710	*CR1*	ENST00000367049	ENSG00000108946	*PRKAR1A*	ENST00000589228
ENSG00000134376	*CRB1*	ENST00000367400	ENSG00000188191	*PRKAR1B*	ENST00000406797
ENSG00000137504	*CREBZF*	ENST00000527447	ENSG00000116132	*PRRX1*	ENST00000239461
ENSG00000204414	*CSHL1*	ENST00000309894	ENSG00000204983	*PRSS1*	ENST00000311737
ENSG00000102974	*CTCF*	ENST00000264010	ENSG00000108671	*PSMD11*	ENST00000261712
ENSG00000168036	*CTNNB1*	ENST00000349496	ENSG00000185920	*PTCH1*	ENST00000331920
ENSG00000158290	*CUL4B*	ENST00000404115	ENSG00000171862	*PTEN*	ENST00000371953
ENSG00000083799	*CYLD*	ENST00000427738	ENSG00000179295	*PTPN11*	ENST00000351677
ENSG00000126733	*DACH2*	ENST00000373125	ENSG00000163348	*PYGO2*	ENST00000368457
ENSG00000189186	*DCAF8L2*	ENST00000451261	ENSG00000112531	*QKI*	ENST00000361752
ENSG00000100523	*DDHD1*	ENST00000323669	ENSG00000136238	*RAC1*	ENST00000356142
ENSG00000100697	*DICER1*	ENST00000526495	ENSG00000185379	*RAD51D*	ENST00000590016
ENSG00000130826	*DKC1*	ENST00000369550	ENSG00000145715	*RASA1*	ENST00000274376
ENSG00000186047	*DLEU7*	ENST00000400393	ENSG00000105538	*RASIP1*	ENST00000222145
ENSG00000137090	*DMRT1*	ENST00000382276	ENSG00000139687	*RB1*	ENST00000267163
ENSG00000187957	*DNER*	ENST00000341772	ENSG00000182872	*RBM10*	ENST00000377604
ENSG00000130816	*DNMT1*	ENST00000359526	ENSG00000173933	*RBM4*	ENST00000409406
ENSG00000119772	*DNMT3A*	ENST00000264709	ENSG00000163694	*RBM47*	ENST00000381793
ENSG00000175920	*DOK7*	ENST00000340083	ENSG00000147274	*RBMX*	ENST00000320676
ENSG00000167130	*DOLPP1*	ENST00000372546	ENSG00000168214	*RBPJ*	ENST00000342295
ENSG00000167261	*DPEP2*	ENST00000412757	ENSG00000166965	*RCCD1*	ENST00000394258
ENSG00000152591	*DSPP*	ENST00000399271	ENSG00000165731	*RET*	ENST00000355710
ENSG00000146648	*EGFR*	ENST00000275493	ENSG00000223638	*RFPL4A*	ENST00000434937

ENSG00000197561	*ELANE*	ENST00000590230	ENSG00000132005	*RFX1*	ENST00000254325
ENSG00000163435	*ELF3*	ENST00000359651	ENSG00000174136	*RGMB*	ENST00000308234
ENSG00000126749	*EMG1*	ENST00000261406	ENSG00000169629	*RGPD8*	ENST00000302558
ENSG00000163508	*EOMES*	ENST00000295743	ENSG00000132677	*RHBG*	ENST00000368249
ENSG00000100393	*EP300*	ENST00000263253	ENSG00000067560	*RHOA*	ENST00000418115
ENSG00000183495	*EP400*	ENST00000389561	ENSG00000136104	*RNASEH2B*	ENST00000336617
ENSG00000116016	*EPAS1*	ENST00000263734	ENSG00000181481	*RNF135*	ENST00000328381
ENSG00000086289	*EPDR1*	ENST00000199448	ENSG00000189051	*RNF222*	ENST00000399398
ENSG00000141736	*ERBB2*	ENST00000269571	ENSG00000204618	*RNF39*	ENST00000244360
ENSG00000187017	*ESPN*	ENST00000377828	ENSG00000156313	*RPGR*	ENST00000378505
ENSG00000196482	*ESRRG*	ENST00000366937	ENSG00000165496	*RPL10L*	ENST00000298283
ENSG00000182197	*EXT1*	ENST00000378204	ENSG00000122406	*RPL5*	ENST00000370321
ENSG00000151348	*EXT2*	ENST00000395673	ENSG00000117676	*RPS6KA1*	ENST00000531382
ENSG00000188107	*EYS*	ENST00000503581	ENSG00000144580	*RQCD1*	ENST00000273064
ENSG00000106462	*EZH2*	ENST00000320356	ENSG00000124782	*RREB1*	ENST00000379938
ENSG00000198734	*F5*	ENST00000367797	ENSG00000159216	*RUNX1*	ENST00000300305
ENSG00000103876	*FAH*	ENST00000407106	ENSG00000124813	*RUNX2*	ENST00000371438
ENSG00000183688	*FAM101B*	ENST00000329099	ENSG00000186350	*RXRA*	ENST00000481739
ENSG00000184731	*FAM110C*	ENST00000327669	ENSG00000119042	*SATB2*	ENST00000417098
ENSG00000147724	*FAM135B*	ENST00000395297	ENSG00000185313	*SCN10A*	ENST00000449082
ENSG00000182230	*FAM153B*	ENST00000515817	ENSG00000168356	*SCN11A*	ENST00000302328
ENSG00000183807	*FAM162B*	ENST00000368557	ENSG00000170616	*SCRT1*	ENST00000332135
ENSG00000185442	*FAM174B*	ENST00000327355	ENSG00000167985	*SDHAF2*	ENST00000301761
ENSG00000047662	*FAM184B*	ENST00000265018	ENSG00000117118	*SDHB*	ENST00000375499
ENSG00000165837	*FAM194B*	ENST00000298738	ENSG00000204370	*SDHD*	ENST00000375549
ENSG00000183508	*FAM46C*	ENST00000369448	ENSG00000255292	*SDHD*	ENST00000532699
ENSG00000174016	*FAM46D*	ENST00000538312	ENSG00000197249	*SERPINA1*	ENST00000448921
ENSG00000188610	*FAM72B*	ENST00000369390	ENSG00000057149	*SERPINB3*	ENST00000283752
ENSG00000180921	*FAM83H*	ENST00000388913	ENSG00000139718	*SETD1B*	ENST00000267197
ENSG00000145002	*FAM86B2*	ENST00000262365	ENSG00000181555	*SETD2*	ENST00000409792
ENSG00000187741	*FANCA*	ENST00000389301	ENSG00000168066	*SF1*	ENST00000377387
ENSG00000181544	*FANCB*	ENST00000398334	ENSG00000115524	*SF3B1*	ENST00000335508
ENSG00000183161	*FANCF*	ENST00000327470	ENSG00000158352	*SHROOM4*	ENST00000376020
ENSG00000026103	*FAS*	ENST00000355740	ENSG00000090402	*SI*	ENST00000264382
ENSG00000165323	*FAT3*	ENST00000298047	ENSG00000254415	*SIGLEC14*	ENST00000360844
ENSG00000112787	*FBRSL1*	ENST00000434748	ENSG00000184302	*SIX6*	ENST00000327720
ENSG00000109670	*FBXW7*	ENST00000281708	ENSG00000157933	*SKI*	ENST00000378536
ENSG00000146618	*FERD3L*	ENST00000275461	ENSG00000091137	*SLC26A4*	ENST00000265715
ENSG00000066468	*FGFR2*	ENST00000457416	ENSG00000139209	*SLC38A4*	ENST00000447411
ENSG00000068078	*FGFR3*	ENST00000340107	ENSG00000188687	*SLC4A5*	ENST00000377634
ENSG00000091483	*FH*	ENST00000366560	ENSG00000188827	*SLX4*	ENST00000294008
ENSG00000154803	*FLCN*	ENST00000285071	ENSG00000175387	*SMAD2*	ENST00000402690
ENSG00000136068	*FLNB*	ENST00000490882	ENSG00000141646	*SMAD4*	ENST00000342988
ENSG00000122025	*FLT3*	ENST00000241453	ENSG00000127616	*SMARCA4*	ENST00000429416
ENSG00000129514	*FOXA1*	ENST00000250448	ENSG00000099956	*SMARCB1*	ENST00000263121
ENSG00000125798	*FOXA2*	ENST00000419308	ENSG00000073584	*SMARCE1*	ENST00000348513
ENSG00000184492	*FOXD4L1*	ENST00000306507	ENSG00000072501	*SMC1A*	ENST00000322213
ENSG00000178919	*FOXE1*	ENST00000375123	ENSG00000108055	*SMC3*	ENST00000361804
ENSG00000164379	*FOXQ1*	ENST00000296839	ENSG00000188176	*SMTNL2*	ENST00000389313
ENSG00000151474	*FRMD4A*	ENST00000357447	ENSG00000132639	*SNAP25*	ENST00000254976
ENSG00000167996	*FTH1*	ENST00000273550	ENSG00000162804	*SNED1*	ENST00000310397
ENSG00000162613	*FUBP1*	ENST00000370768	ENSG00000100028	*SNRPD3*	ENST00000215829
ENSG00000157240	*FZD1*	ENST00000287934	ENSG00000115904	*SOS1*	ENST00000426016
ENSG00000109158	*GABRA4*	ENST00000264318	ENSG00000164736	*SOX17*	ENST00000297316
ENSG00000113327	*GABRG2*	ENST00000414552	ENSG00000124766	*SOX4*	ENST00000244745
ENSG00000179348	*GATA2*	ENST00000341105	ENSG00000125398	*SOX9*	ENST00000245479

ENSG00000107485	*GATA3*	ENST00000379328	ENSG00000164651	*SP8*	ENST00000418710
ENSG00000177628	*GBA*	ENST00000327247	ENSG00000203923	*SPANXN1*	ENST00000370493
ENSG00000100116	*GCAT*	ENST00000323205	ENSG00000133104	*SPG20*	ENST00000451493
ENSG00000123159	*GIPC1*	ENST00000393033	ENSG00000147059	*SPIN2A*	ENST00000374908
ENSG00000165474	*GJB2*	ENST00000382844	ENSG00000121067	*SPOP*	ENST00000393331
ENSG00000106571	*GLI3*	ENST00000395925	ENSG00000167978	*SRRM2*	ENST00000301740
ENSG00000088256	*GNA11*	ENST00000078429	ENSG00000157216	*SSBP3*	ENST00000371320
ENSG00000156052	*GNAQ*	ENST00000286548	ENSG00000101972	*STAG2*	ENST00000218089
ENSG00000087460	*GNAS*	ENST00000371100	ENSG00000168610	*STAT3*	ENST00000264657
ENSG00000172380	*GNG12*	ENST00000370982	ENSG00000118046	*STK11*	ENST00000326873
ENSG00000215405	*GOLGA6L6*	ENST00000427390	ENSG00000107882	*SUFU*	ENST00000369902
ENSG00000147257	*GPC3*	ENST00000394299	ENSG00000122012	*SV2C*	ENST00000502798
ENSG00000146360	*GPR6*	ENST00000275169	ENSG00000176438	*SYNE3*	ENST00000334258
ENSG00000204175	*GPRIN2*	ENST00000374314	ENSG00000148835	*TAF5*	ENST00000369839
ENSG00000109519	*GRPEL1*	ENST00000264954	ENSG00000122145	*TBX22*	ENST00000373294
ENSG00000148702	*HABP2*	ENST00000351270	ENSG00000135111	*TBX3*	ENST00000257566
ENSG00000223609	*HBD*	ENST00000380299	ENSG00000121075	*TBX4*	ENST00000240335
ENSG00000196565	*HBG2*	ENST00000380259	ENSG00000204065	*TCEAL5*	ENST00000372680
ENSG00000002746	*HECW1*	ENST00000395891	ENSG00000148737	*TCF7L2*	ENST00000543371
ENSG00000182218	*HHIPL1*	ENST00000330710	ENSG00000164362	*TERT*	ENST00000310581
ENSG00000256269	*HMBS*	ENST00000278715	ENSG00000106799	*TGFBR1*	ENST00000374994
ENSG00000135100	*HNF1A*	ENST00000257555	ENSG00000163513	*TGFBR2*	ENST00000359013
ENSG00000165119	*HNRNPK*	ENST00000376263	ENSG00000153779	*TGIF2LX*	ENST00000561129
ENSG00000174775	*HRAS*	ENST00000451590	ENSG00000144229	*THSD7B*	ENST00000272643
ENSG00000196196	*HRCT1*	ENST00000354323	ENSG00000135956	*TMEM127*	ENST00000258439
ENSG00000108786	*HSD17B1*	ENST00000585807	ENSG00000206432	*TMEM200C*	ENST00000581347
ENSG00000102878	*HSF4*	ENST00000264009	ENSG00000234224	*TMEM229A*	ENST00000455783
ENSG00000127415	*IDUA*	ENST00000247933	ENSG00000157873	*TNFRSF14*	ENST00000355716
ENSG00000162783	*IER5*	ENST00000367577	ENSG00000106952	*TNFSF8*	ENST00000223795
ENSG00000142089	*IFITM3*	ENST00000399808	ENSG00000141510	*TP53*	ENST00000269305
ENSG00000254709	*IGLL5*	ENST00000526893	ENSG00000095917	*TPSD1*	ENST00000211076
ENSG00000168310	*IRF2*	ENST00000393593	ENSG00000166157	*TPTE*	ENST00000361285
ENSG00000137265	*IRF4*	ENST00000380956	ENSG00000132958	*TPTE2*	ENST00000400230
ENSG00000133124	*IRS4*	ENST00000372129	ENSG00000131323	*TRAF3*	ENST00000560371
ENSG00000177508	*IRX3*	ENST00000329734	ENSG00000100106	*TRIOBP*	ENST00000406386
ENSG00000152409	*JMY*	ENST00000396137	ENSG00000165699	*TSC1*	ENST00000298552
ENSG00000186994	*KANK3*	ENST00000330915	ENSG00000103197	*TSC2*	ENST00000219476
ENSG00000083168	*KAT6A*	ENST00000396930	ENSG00000196428	*TSC22D2*	ENST00000361875
ENSG00000234438	*KBTBD13*	ENST00000432196	ENSG00000231738	*TSPAN19*	ENST00000532498
ENSG00000180509	*KCNE1*	ENST00000337385	ENSG00000077498	*TYR*	ENST00000263321
ENSG00000164626	*KCNK5*	ENST00000359534	ENSG00000092445	*TYRO3*	ENST00000263798
ENSG00000143603	*KCNN3*	ENST00000271915	ENSG00000160201	*U2AF1*	ENST00000291552
ENSG00000162687	*KCNT2*	ENST00000294725	ENSG00000077721	*UBE2A*	ENST00000371558
ENSG00000126012	*KDM5C*	ENST00000375401	ENSG00000169021	*UQCRFS1*	ENST00000304863
ENSG00000147050	*KDM6A*	ENST00000377967	ENSG00000126088	*UROD*	ENST00000246337
ENSG00000197993	*KEL*	ENST00000355265	ENSG00000181408	*UTS2R*	ENST00000313135
ENSG00000235750	*KIAA0040*	ENST00000545251	ENSG00000177504	*VCX2*	ENST00000317103
ENSG00000157404	*KIT*	ENST00000288135	ENSG00000150630	*VEGFC*	ENST00000280193
ENSG00000109787	*KLF3*	ENST00000261438	ENSG00000134086	*VHL*	ENST00000256474
ENSG00000102554	*KLF5*	ENST00000377687	ENSG00000178201	*VN1R1*	ENST00000321039
ENSG00000055609	*KMT2C*	ENST00000262189	ENSG00000188730	*VWC2*	ENST00000340652
ENSG00000167548	*KMT2D*	ENST00000301067	ENSG00000015285	*WAS*	ENST00000376701
ENSG00000133703	*KRAS*	ENST00000256078	ENSG00000184937	*WT1*	ENST00000332351
ENSG00000171346	*KRT15*	ENST00000254043	ENSG00000181722	*ZBTB20*	ENST00000474710
ENSG00000172867	*KRT2*	ENST00000309680	ENSG00000160685	*ZBTB7B*	ENST00000417934
ENSG00000139648	*KRT71*	ENST00000267119	ENSG00000178199	*ZC3H12D*	ENST00000409806

ENSG00000161849	*KRT84*	ENST00000257951	ENSG00000177764	*ZCCHC3*	ENST00000382352
ENSG00000212721	*KRTAP4-11*	ENST00000391413	ENSG00000156599	*ZDHHC5*	ENST00000287169
ENSG00000240871	*KRTAP4-7*	ENST00000391417	ENSG00000169554	*ZEB2*	ENST00000558170
ENSG00000107929	*LARP4B*	ENST00000316157	ENSG00000140836	*ZFHX3*	ENST00000268489
ENSG00000169744	*LDB2*	ENST00000304523	ENSG00000152518	*ZFP36L2*	ENST00000282388
ENSG00000168924	*LETM1*	ENST00000302787	ENSG00000179588	*ZFPM1*	ENST00000319555
ENSG00000182541	*LIMK2*	ENST00000340552	ENSG00000122515	*ZMIZ2*	ENST00000309315
ENSG00000101670	*LIPG*	ENST00000261292	ENSG00000267508	*ZNF285*	ENST00000330997
ENSG00000164715	*LMTK2*	ENST00000297293	ENSG00000167962	*ZNF598*	ENST00000431526
ENSG00000203782	*LOR*	ENST00000368742	ENSG00000197123	*ZNF679*	ENST00000421025
ENSG00000120256	*LRP11*	ENST00000239367	ENSG00000196456	*ZNF775*	ENST00000329630
ENSG00000131409	*LRRC4B*	ENST00000599957	ENSG00000204514	*ZNF814*	ENST00000435989
ENSG00000148948	*LRRC4C*	ENST00000278198	ENSG00000234284	*ZNF879*	ENST00000444149
ENSG00000125872	*LRRN4*	ENST00000378858	ENSG00000213973	*ZNF99*	ENST00000596209
ENSG00000168056	*LTBP3*	ENST00000301873	ENSG00000188372	*ZP3*	ENST00000394857
ENSG00000062524	*LTK*	ENST00000263800			

Table A5 - Gene list used for analysis of truncating variants based on somatically mutated genes and cancer known CPGs – Refined with WebGestalt

Gene identifier	Gene name	Canonical transcript	Gene identifier	Gene name	Canonical transcript
ENSG00000115977	*AAK1*	ENST00000409085	ENSG00000133020	*MYH8*	ENST00000403437
ENSG00000085563	*ABCB1*	ENST00000265724	ENSG00000128641	*MYO1B*	ENST00000392318
ENSG00000069431	*ABCC9*	ENST00000261200	ENSG00000186462	*NAP1L2*	ENST00000373517
ENSG00000123983	*ACSL3*	ENST00000357430	ENSG00000144035	*NAT8*	ENST00000272425
ENSG00000075624	*ACTB*	ENST00000331789	ENSG00000104320	*NBN*	ENST00000265433
ENSG00000184009	*ACTG1*	ENST00000575842	ENSG00000158092	*NCK1*	ENST00000481752
ENSG00000077080	*ACTL6B*	ENST00000160382	ENSG00000124151	*NCOA3*	ENST00000371998
ENSG00000135503	*ACVR1B*	ENST00000541224	ENSG00000141027	*NCOR1*	ENST00000268712
ENSG00000145536	*ADAMTS16*	ENST00000274181	ENSG00000129559	*NEDD8*	ENST00000250495
ENSG00000087116	*ADAMTS2*	ENST00000251582	ENSG00000100285	*NEFH*	ENST00000310624
ENSG00000106624	*AEBP1*	ENST00000223357	ENSG00000196712	*NF1*	ENST00000358273
ENSG00000144891	*AGTR1*	ENST00000542281	ENSG00000186575	*NF2*	ENST00000338641
ENSG00000113492	*AGXT2*	ENST00000231420	ENSG00000116044	*NFE2L2*	ENST00000397062
ENSG00000110711	*AIP*	ENST00000279146	ENSG00000164190	*NIPBL*	ENST00000282516
ENSG00000129474	*AJUBA*	ENST00000262713	ENSG00000167034	*NKX3-1*	ENST00000380871
ENSG00000151320	*AKAP6*	ENST00000280979	ENSG00000087095	*NLK*	ENST00000407008
ENSG00000142208	*AKT1*	ENST00000554581	ENSG00000171487	*NLRP5*	ENST00000390649
ENSG00000171094	*ALK*	ENST00000389048	ENSG00000174885	*NLRP6*	ENST00000312165
ENSG00000184675	*AMER1*	ENST00000330258	ENSG00000197696	*NMB*	ENST00000394588
ENSG00000135409	*AMHR2*	ENST00000257863	ENSG00000169251	*NMD3*	ENST00000460469
ENSG00000130812	*ANGPTL6*	ENST00000253109	ENSG00000109255	*NMU*	ENST00000264218
ENSG00000148513	*ANKRD30A*	ENST00000361713	ENSG00000148400	*NOTCH1*	ENST00000277541
ENSG00000134982	*APC*	ENST00000457016	ENSG00000134250	*NOTCH2*	ENST00000256646
ENSG00000132703	*APCS*	ENST00000255040	ENSG00000188747	*NOXA1*	ENST00000341349
ENSG00000130203	*APOE*	ENST00000252486	ENSG00000056291	*NPFFR2*	ENST00000308744
ENSG00000178878	*APOLD1*	ENST00000326765	ENSG00000181163	*NPM1*	ENST00000296930
ENSG00000169083	*AR*	ENST00000374690	ENSG00000171246	*NPTX1*	ENST00000306773
ENSG00000117713	*ARID1A*	ENST00000324856	ENSG00000213281	*NRAS*	ENST00000369535
ENSG00000189079	*ARID2*	ENST00000334344	ENSG00000106459	*NRF1*	ENST00000393232
ENSG00000163466	*ARPC2*	ENST00000295685	ENSG00000123572	*NRK*	ENST00000428173
ENSG00000161664	*ASB16*	ENST00000293414	ENSG00000165671	*NSD1*	ENST00000439151
ENSG00000187855	*ASCL4*	ENST00000342331	ENSG00000065057	*NTHL1*	ENST00000219066
ENSG00000148219	*ASTN2*	ENST00000361209	ENSG00000074590	*NUAK1*	ENST00000261402
ENSG00000143970	*ASXL2*	ENST00000435504	ENSG00000105245	*NUMBL*	ENST00000252891
ENSG00000149311	*ATM*	ENST00000278616	ENSG00000102900	*NUP93*	ENST00000308159
ENSG00000111676	*ATN1*	ENST00000356654	ENSG00000154358	*OBSCN*	ENST00000570156
ENSG00000168874	*ATOH8*	ENST00000306279	ENSG00000087263	*OGFOD1*	ENST00000566157
ENSG00000116039	*ATP6V1B1*	ENST00000234396	ENSG00000130558	*OLFM1*	ENST00000252854
ENSG00000085224	*ATRX*	ENST00000373344	ENSG00000116329	*OPRD1*	ENST00000234961
ENSG00000124788	*ATXN1*	ENST00000244769	ENSG00000234560	*OR10G8*	ENST00000431524
ENSG00000066427	*ATXN3*	ENST00000393287	ENSG00000257019	*OR13C2*	ENST00000542196
ENSG00000103126	*AXIN1*	ENST00000262320	ENSG00000172150	*OR1A2*	ENST00000381951
ENSG00000168646	*AXIN2*	ENST00000307078	ENSG00000197887	*OR1S2*	ENST00000302592
ENSG00000166710	*B2M*	ENST00000558401	ENSG00000221938	*OR2A14*	ENST00000408899
ENSG00000175866	*BAIAP2*	ENST00000321300	ENSG00000221989	*OR2A2*	ENST00000408979
ENSG00000163930	*BAP1*	ENST00000460680	ENSG00000188558	*OR2G6*	ENST00000343414
ENSG00000127152	*BCL11B*	ENST00000357195	ENSG00000196071	*OR2L13*	ENST00000366478
ENSG00000110987	*BCL7A*	ENST00000538010	ENSG00000196936	*OR2L8*	ENST00000357191
ENSG00000180828	*BHLHE22*	ENST00000321870	ENSG00000162727	*OR2M5*	ENST00000366476
ENSG00000197299	*BLM*	ENST00000355112	ENSG00000177201	*OR2T12*	ENST00000317996

ENSG00000183682	*BMP8A*	ENST00000331593	ENSG00000196240	*OR2T2*	ENST00000342927
ENSG00000107779	*BMPR1A*	ENST00000372037	ENSG00000177212	*OR2T33*	ENST00000318021
ENSG00000157764	*BRAF*	ENST00000288602	ENSG00000183310	*OR2T34*	ENST00000328782
ENSG00000012048	*BRCA1*	ENST00000471181	ENSG00000196944	*OR2T4*	ENST00000366475
ENSG00000139618	*BRCA2*	ENST00000544455	ENSG00000177462	*OR2T8*	ENST00000319968
ENSG00000112983	*BRD8*	ENST00000254900	ENSG00000221840	*OR4A5*	ENST00000319760
ENSG00000162670	*BRINP3*	ENST00000367462	ENSG00000181935	*OR4C16*	ENST00000314634
ENSG00000136492	*BRIP1*	ENST00000259008	ENSG00000176547	*OR4C3*	ENST00000319856
ENSG00000151136	*BTBD11*	ENST00000280758	ENSG00000141194	*OR4D1*	ENST00000268912
ENSG00000159388	*BTG2*	ENST00000290551	ENSG00000176200	*OR4D11*	ENST00000313253
ENSG00000156970	*BUB1B*	ENST00000287598	ENSG00000182854	*OR4F15*	ENST00000332238
ENSG00000039537	*C6*	ENST00000263413	ENSG00000182974	*OR4M2*	ENST00000332663
ENSG00000157131	*C8A*	ENST00000361249	ENSG00000176294	*OR4N2*	ENST00000315947
ENSG00000004948	*CALCR*	ENST00000359558	ENSG00000176895	*OR51A7*	ENST00000359350
ENSG00000108509	*CAMTA2*	ENST00000414043	ENSG00000184881	*OR51B2*	ENST00000328813
ENSG00000064012	*CASP8*	ENST00000358485	ENSG00000242180	*OR51B5*	ENST00000300773
ENSG00000118729	*CASQ2*	ENST00000261448	ENSG00000176879	*OR51G1*	ENST00000321961
ENSG00000067955	*CBFB*	ENST00000412916	ENSG00000176893	*OR51G2*	ENST00000322013
ENSG00000110395	*CBL*	ENST00000264033	ENSG00000181609	*OR52D1*	ENST00000322641
ENSG00000122565	*CBX3*	ENST00000337620	ENSG00000176937	*OR52R1*	ENST00000356069
ENSG00000128596	*CCDC136*	ENST00000297788	ENSG00000172459	*OR5AR1*	ENST00000302969
ENSG00000180376	*CCDC66*	ENST00000394672	ENSG00000198877	*OR5D13*	ENST00000361760
ENSG00000106178	*CCL24*	ENST00000416943	ENSG00000149133	*OR5F1*	ENST00000278409
ENSG00000110092	*CCND1*	ENST00000227507	ENSG00000231192	*OR5H1*	ENST00000354565
ENSG00000010610	*CD4*	ENST00000011653	ENSG00000236032	*OR5H14*	ENST00000437310
ENSG00000114013	*CD86*	ENST00000330540	ENSG00000233412	*OR5H15*	ENST00000356526
ENSG00000004897	*CDC27*	ENST00000531206	ENSG00000186117	*OR5L1*	ENST00000333973
ENSG00000134371	*CDC73*	ENST00000367435	ENSG00000205030	*OR5L2*	ENST00000378397
ENSG00000039068	*CDH1*	ENST00000261769	ENSG00000174937	*OR5M3*	ENST00000312240
ENSG00000167258	*CDK12*	ENST00000447079	ENSG00000174942	*OR5R1*	ENST00000312253
ENSG00000135446	*CDK4*	ENST00000257904	ENSG00000172489	*OR5T3*	ENST00000303059
ENSG00000124762	*CDKN1A*	ENST00000405375	ENSG00000187612	*OR5W2*	ENST00000344514
ENSG00000111276	*CDKN1B*	ENST00000228872	ENSG00000169214	*OR6F1*	ENST00000302084
ENSG00000129757	*CDKN1C*	ENST00000414822	ENSG00000203757	*OR6K3*	ENST00000368145
ENSG00000147889	*CDKN2A*	ENST00000498124	ENSG00000198657	*OR8B4*	ENST00000356130
ENSG00000147883	*CDKN2B*	ENST00000276925	ENSG00000172154	*OR8I2*	ENST00000302124
ENSG00000245848	*CEBPA*	ENST00000498907	ENSG00000181689	*OR8K3*	ENST00000312711
ENSG00000139610	*CELA1*	ENST00000293636	ENSG00000181752	*OR8K5*	ENST00000313447
ENSG00000166037	*CEP57*	ENST00000325542	ENSG00000197376	*OR8S1*	ENST00000310194
ENSG00000111642	*CHD4*	ENST00000357008	ENSG00000165588	*OTX2*	ENST00000339475
ENSG00000183765	*CHEK2*	ENST00000382580	ENSG00000155463	*OXA1L*	ENST00000285848
ENSG00000131873	*CHSY1*	ENST00000254190	ENSG00000182162	*P2RY8*	ENST00000381297
ENSG00000079432	*CIC*	ENST00000575354	ENSG00000070756	*PABPC1*	ENST00000318607
ENSG00000174600	*CMKLR1*	ENST00000312143	ENSG00000083093	*PALB2*	ENST00000261584
ENSG00000060718	*COL11A1*	ENST00000370096	ENSG00000125779	*PANK2*	ENST00000316562
ENSG00000164692	*COL1A2*	ENST00000297268	ENSG00000162073	*PAQR4*	ENST00000318782
ENSG00000114270	*COL7A1*	ENST00000328333	ENSG00000007372	*PAX6*	ENST00000419022
ENSG00000021826	*CPS1*	ENST00000430249	ENSG00000163939	*PBRM1*	ENST00000394830
ENSG00000203710	*CR1*	ENST00000367049	ENSG00000165494	*PCF11*	ENST00000298281
ENSG00000134376	*CRB1*	ENST00000367400	ENSG00000056661	*PCGF2*	ENST00000580830
ENSG00000137504	*CREBZF*	ENST00000527447	ENSG00000156374	*PCGF6*	ENST00000369847
ENSG00000213145	*CRIP1*	ENST00000330233	ENSG00000249915	*PDCD6*	ENST00000264933
ENSG00000121552	*CSTA*	ENST00000264474	ENSG00000134853	*PDGFRA*	ENST00000257290
ENSG00000102974	*CTCF*	ENST00000264010	ENSG00000049246	*PER3*	ENST00000361923
ENSG00000118523	*CTGF*	ENST00000367976	ENSG00000082175	*PGR*	ENST00000325455
ENSG00000168036	*CTNNB1*	ENST00000349496	ENSG00000164040	*PGRMC2*	ENST00000520121

ENSG00000158290	*CUL4B*	ENST00000404115	ENSG00000156531	*PHF6*	ENST00000332070
ENSG00000083799	*CYLD*	ENST00000427738	ENSG00000109132	*PHOX2B*	ENST00000226382
ENSG00000172817	*CYP7B1*	ENST00000310193	ENSG00000105229	*PIAS4*	ENST00000262971
ENSG00000152207	*CYSLTR2*	ENST00000282018	ENSG00000121879	*PIK3CA*	ENST00000263967
ENSG00000126733	*DACH2*	ENST00000373125	ENSG00000145675	*PIK3R1*	ENST00000521381
ENSG00000133083	*DCLK1*	ENST00000255448	ENSG00000170890	*PLA2G1B*	ENST00000308366
ENSG00000134574	*DDB2*	ENST00000256996	ENSG00000214456	*PLIN5*	ENST00000381848
ENSG00000013573	*DDX11*	ENST00000407793	ENSG00000106397	*PLOD3*	ENST00000223127
ENSG00000124795	*DEK*	ENST00000397239	ENSG00000064933	*PMS1*	ENST00000441310
ENSG00000100697	*DICER1*	ENST00000526495	ENSG00000062822	*POLD1*	ENST00000440232
ENSG00000211448	*DIO2*	ENST00000555750	ENSG00000177084	*POLE*	ENST00000320574
ENSG00000144535	*DIS3L2*	ENST00000325385	ENSG00000170734	*POLH*	ENST00000372236
ENSG00000130826	*DKC1*	ENST00000369550	ENSG00000170836	*PPM1D*	ENST00000305921
ENSG00000137090	*DMRT1*	ENST00000382276	ENSG00000105568	*PPP2R1A*	ENST00000322088
ENSG00000187957	*DNER*	ENST00000341772	ENSG00000138814	*PPP3CA*	ENST00000394854
ENSG00000130816	*DNMT1*	ENST00000359526	ENSG00000116721	*PRAMEF1*	ENST00000332296
ENSG00000119772	*DNMT3A*	ENST00000264709	ENSG00000137509	*PRCP*	ENST00000393399
ENSG00000134516	*DOCK2*	ENST00000256935	ENSG00000057657	*PRDM1*	ENST00000369096
ENSG00000107099	*DOCK8*	ENST00000453981	ENSG00000164256	*PRDM9*	ENST00000296682
ENSG00000206052	*DOK6*	ENST00000382713	ENSG00000108946	*PRKAR1A*	ENST00000589228
ENSG00000175920	*DOK7*	ENST00000340083	ENSG00000116132	*PRRX1*	ENST00000239461
ENSG00000121570	*DPPA4*	ENST00000335658	ENSG00000164985	*PSIP1*	ENST00000380733
ENSG00000152591	*DSPP*	ENST00000399271	ENSG00000108671	*PSMD11*	ENST00000261712
ENSG00000112679	*DUSP22*	ENST00000344450	ENSG00000185920	*PTCH1*	ENST00000331920
ENSG00000164176	*EDIL3*	ENST00000296591	ENSG00000171862	*PTEN*	ENST00000371953
ENSG00000136160	*EDNRB*	ENST00000377211	ENSG00000179295	*PTPN11*	ENST00000351677
ENSG00000146648	*EGFR*	ENST00000275493	ENSG00000127947	*PTPN12*	ENST00000248594
ENSG00000120738	*EGR1*	ENST00000239938	ENSG00000163348	*PYGO2*	ENST00000368457
ENSG00000197561	*ELANE*	ENST00000590230	ENSG00000112531	*QKI*	ENST00000361752
ENSG00000163435	*ELF3*	ENST00000359651	ENSG00000136238	*RAC1*	ENST00000356142
ENSG00000155849	*ELMO1*	ENST00000310758	ENSG00000164754	*RAD21*	ENST00000297338
ENSG00000163508	*EOMES*	ENST00000295743	ENSG00000108384	*RAD51C*	ENST00000337432
ENSG00000100393	*EP300*	ENST00000263253	ENSG00000185379	*RAD51D*	ENST00000590016
ENSG00000183495	*EP400*	ENST00000389561	ENSG00000145715	*RASA1*	ENST00000274376
ENSG00000116016	*EPAS1*	ENST00000263734	ENSG00000111344	*RASAL1*	ENST00000546530
ENSG00000119888	*EPCAM*	ENST00000263735	ENSG00000105538	*RASIP1*	ENST00000222145
ENSG00000142627	*EPHA2*	ENST00000358432	ENSG00000139687	*RB1*	ENST00000267163
ENSG00000080224	*EPHA6*	ENST00000389672	ENSG00000182872	*RBM10*	ENST00000377604
ENSG00000141736	*ERBB2*	ENST00000269571	ENSG00000173933	*RBM4*	ENST00000409406
ENSG00000065361	*ERBB3*	ENST00000267101	ENSG00000147274	*RBMX*	ENST00000320676
ENSG00000082805	*ERC1*	ENST00000397203	ENSG00000168214	*RBPJ*	ENST00000342295
ENSG00000104884	*ERCC2*	ENST00000391945	ENSG00000124232	*RBPJL*	ENST00000343694
ENSG00000163161	*ERCC3*	ENST00000285398	ENSG00000160957	*RECQL4*	ENST00000428558
ENSG00000175595	*ERCC4*	ENST00000311895	ENSG00000115386	*REG1A*	ENST00000233735
ENSG00000134899	*ERCC5*	ENST00000355739	ENSG00000172023	*REG1B*	ENST00000305089
ENSG00000187017	*ESPN*	ENST00000377828	ENSG00000172016	*REG3A*	ENST00000393878
ENSG00000196482	*ESRRG*	ENST00000366937	ENSG00000143954	*REG3G*	ENST00000272324
ENSG00000182197	*EXT1*	ENST00000378204	ENSG00000165731	*RET*	ENST00000355710
ENSG00000151348	*EXT2*	ENST00000395673	ENSG00000132005	*RFX1*	ENST00000254325
ENSG00000188107	*EYS*	ENST00000503581	ENSG00000174136	*RGMB*	ENST00000308234
ENSG00000106462	*EZH2*	ENST00000320356	ENSG00000182901	*RGS7*	ENST00000366565
ENSG00000101447	*FAM83D*	ENST00000217429	ENSG00000186326	*RGS9BP*	ENST00000334176
ENSG00000180921	*FAM83H*	ENST00000388913	ENSG00000067560	*RHOA*	ENST00000418115
ENSG00000183304	*FAM9A*	ENST00000543214	ENSG00000143878	*RHOB*	ENST00000272233
ENSG00000187741	*FANCA*	ENST00000389301	ENSG00000119729	*RHOQ*	ENST00000238738
ENSG00000181544	*FANCB*	ENST00000398334	ENSG00000136104	*RNASEH2B*	ENST00000336617

ENSG00000158169	*FANCC*	ENST00000289081	ENSG00000181481	*RNF135*	ENST00000328381
ENSG00000144554	*FANCD2*	ENST00000287647	ENSG00000189051	*RNF222*	ENST00000399398
ENSG00000112039	*FANCE*	ENST00000229769	ENSG00000108375	*RNF43*	ENST00000584437
ENSG00000183161	*FANCF*	ENST00000327470	ENSG00000117676	*RPS6KA1*	ENST00000531382
ENSG00000221829	*FANCG*	ENST00000378643	ENSG00000124782	*RREB1*	ENST00000379938
ENSG00000140525	*FANCI*	ENST00000310775	ENSG00000159216	*RUNX1*	ENST00000300305
ENSG00000115392	*FANCL*	ENST00000402135	ENSG00000079102	*RUNX1T1*	ENST00000436581
ENSG00000187790	*FANCM*	ENST00000267430	ENSG00000124813	*RUNX2*	ENST00000371438
ENSG00000026103	*FAS*	ENST00000355740	ENSG00000186350	*RXRA*	ENST00000481739
ENSG00000109670	*FBXW7*	ENST00000281708	ENSG00000119042	*SATB2*	ENST00000417098
ENSG00000203747	*FCGR3A*	ENST00000367969	ENSG00000126524	*SBDS*	ENST00000246868
ENSG00000146618	*FERD3L*	ENST00000275461	ENSG00000170616	*SCRT1*	ENST00000332135
ENSG00000171055	*FEZ2*	ENST00000379245	ENSG00000137575	*SDCBP*	ENST00000260130
ENSG00000066468	*FGFR2*	ENST00000457416	ENSG00000167985	*SDHAF2*	ENST00000301761
ENSG00000068078	*FGFR3*	ENST00000340107	ENSG00000007908	*SELE*	ENST00000333360
ENSG00000134775	*FHOD3*	ENST00000257209	ENSG00000075223	*SEMA3C*	ENST00000265361
ENSG00000154803	*FLCN*	ENST00000285071	ENSG00000082684	*SEMA5B*	ENST00000451055
ENSG00000143631	*FLG*	ENST00000368799	ENSG00000057149	*SERPINB3*	ENST00000283752
ENSG00000136068	*FLNB*	ENST00000490882	ENSG00000181555	*SETD2*	ENST00000409792
ENSG00000122025	*FLT3*	ENST00000241453	ENSG00000168066	*SF1*	ENST00000377387
ENSG00000129514	*FOXA1*	ENST00000250448	ENSG00000115524	*SF3B1*	ENST00000335508
ENSG00000125798	*FOXA2*	ENST00000419308	ENSG00000118515	*SGK1*	ENST00000367858
ENSG00000184492	*FOXD4L1*	ENST00000306507	ENSG00000158352	*SHROOM4*	ENST00000376020
ENSG00000184659	*FOXD4L4*	ENST00000377413	ENSG00000112246	*SIM1*	ENST00000369208
ENSG00000178919	*FOXE1*	ENST00000375123	ENSG00000198053	*SIRPA*	ENST00000358771
ENSG00000164379	*FOXQ1*	ENST00000296839	ENSG00000184302	*SIX6*	ENST00000327720
ENSG00000136877	*FPGS*	ENST00000373247	ENSG00000157933	*SKI*	ENST00000378536
ENSG00000109536	*FRG1*	ENST00000226798	ENSG00000014824	*SLC30A9*	ENST00000264451
ENSG00000167996	*FTH1*	ENST00000273550	ENSG00000148482	*SLC39A12*	ENST00000377369
ENSG00000162613	*FUBP1*	ENST00000370768	ENSG00000188687	*SLC4A5*	ENST00000377634
ENSG00000157240	*FZD1*	ENST00000287934	ENSG00000184564	*SLITRK6*	ENST00000400286
ENSG00000104290	*FZD3*	ENST00000240093	ENSG00000188827	*SLX4*	ENST00000294008
ENSG00000109158	*GABRA4*	ENST00000264318	ENSG00000175387	*SMAD2*	ENST00000402690
ENSG00000145863	*GABRA6*	ENST00000274545	ENSG00000141646	*SMAD4*	ENST00000342988
ENSG00000113327	*GABRG2*	ENST00000414552	ENSG00000127616	*SMARCA4*	ENST00000429416
ENSG00000146276	*GABRR1*	ENST00000454853	ENSG00000099956	*SMARCB1*	ENST00000263121
ENSG00000179348	*GATA2*	ENST00000341105	ENSG00000073584	*SMARCE1*	ENST00000348513
ENSG00000107485	*GATA3*	ENST00000379328	ENSG00000072501	*SMC1A*	ENST00000322213
ENSG00000177628	*GBA*	ENST00000327247	ENSG00000108055	*SMC3*	ENST00000361804
ENSG00000179168	*GGN*	ENST00000334928	ENSG00000116698	*SMG7*	ENST00000507469
ENSG00000123159	*GIPC1*	ENST00000393033	ENSG00000132639	*SNAP25*	ENST00000254976
ENSG00000165474	*GJB2*	ENST00000382844	ENSG00000104976	*SNAPC2*	ENST00000221573
ENSG00000106571	*GLI3*	ENST00000395925	ENSG00000100028	*SNRPD3*	ENST00000215829
ENSG00000104499	*GML*	ENST00000220940	ENSG00000115904	*SOS1*	ENST00000426016
ENSG00000088256	*GNA11*	ENST00000078429	ENSG00000164736	*SOX17*	ENST00000297316
ENSG00000156052	*GNAQ*	ENST00000286548	ENSG00000124766	*SOX4*	ENST00000244745
ENSG00000087460	*GNAS*	ENST00000371100	ENSG00000125398	*SOX9*	ENST00000245479
ENSG00000172380	*GNG12*	ENST00000370982	ENSG00000105866	*SP4*	ENST00000222584
ENSG00000147257	*GPC3*	ENST00000394299	ENSG00000164651	*SP8*	ENST00000418710
ENSG00000183098	*GPC6*	ENST00000377047	ENSG00000163071	*SPATA18*	ENST00000295213
ENSG00000146360	*GPR6*	ENST00000275169	ENSG00000141255	*SPATA22*	ENST00000573128
ENSG00000132522	*GPS2*	ENST00000380728	ENSG00000133104	*SPG20*	ENST00000451493
ENSG00000215203	*GRXCR1*	ENST00000399770	ENSG00000121067	*SPOP*	ENST00000393331
ENSG00000244067	*GSTA2*	ENST00000493422	ENSG00000169474	*SPRR1A*	ENST00000307122
ENSG00000010704	*HFE*	ENST00000357618	ENSG00000163554	*SPTA1*	ENST00000368147
ENSG00000182218	*HHIPL1*	ENST00000330710	ENSG00000167978	*SRRM2*	ENST00000301740

ENSG00000114455	*HHLA2*	ENST00000357759	ENSG00000184895	*SRY*	ENST00000383070
ENSG00000168298	*HIST1H1E*	ENST00000304218	ENSG00000157216	*SSBP3*	ENST00000371320
ENSG00000164508	*HIST1H2AA*	ENST00000297012	ENSG00000101972	*STAG2*	ENST00000218089
ENSG00000205581	*HMGN1*	ENST00000380749	ENSG00000168610	*STAT3*	ENST00000264657
ENSG00000135100	*HNF1A*	ENST00000257555	ENSG00000118046	*STK11*	ENST00000326873
ENSG00000165119	*HNRNPK*	ENST00000376263	ENSG00000107882	*SUFU*	ENST00000369902
ENSG00000099783	*HNRNPM*	ENST00000325495	ENSG00000173597	*SULT1B1*	ENST00000310613
ENSG00000174775	*HRAS*	ENST00000451590	ENSG00000178691	*SUZ12*	ENST00000322652
ENSG00000102878	*HSF4*	ENST00000264009	ENSG00000131018	*SYNE1*	ENST00000367255
ENSG00000138413	*IDH1*	ENST00000415913	ENSG00000176438	*SYNE3*	ENST00000334258
ENSG00000182054	*IDH2*	ENST00000330062	ENSG00000148835	*TAF5*	ENST00000369839
ENSG00000142089	*IFITM3*	ENST00000399808	ENSG00000169777	*TAS2R1*	ENST00000382492
ENSG00000136634	*IL10*	ENST00000423557	ENSG00000121318	*TAS2R10*	ENST00000240619
ENSG00000115607	*IL18RAP*	ENST00000264260	ENSG00000127362	*TAS2R3*	ENST00000247879
ENSG00000016402	*IL20RA*	ENST00000316649	ENSG00000255374	*TAS2R43*	ENST00000531678
ENSG00000134352	*IL6ST*	ENST00000381298	ENSG00000122145	*TBX22*	ENST00000373294
ENSG00000153487	*ING1*	ENST00000375774	ENSG00000135111	*TBX3*	ENST00000257566
ENSG00000168310	*IRF2*	ENST00000393593	ENSG00000121075	*TBX4*	ENST00000240335
ENSG00000137265	*IRF4*	ENST00000380956	ENSG00000204065	*TCEAL5*	ENST00000372680
ENSG00000133124	*IRS4*	ENST00000372129	ENSG00000113649	*TCERG1*	ENST00000296702
ENSG00000177508	*IRX3*	ENST00000329734	ENSG00000140262	*TCF12*	ENST00000438423
ENSG00000113263	*ITK*	ENST00000422843	ENSG00000148737	*TCF7L2*	ENST00000543371
ENSG00000152409	*JMY*	ENST00000396137	ENSG00000164362	*TERT*	ENST00000310581
ENSG00000083168	*KAT6A*	ENST00000396930	ENSG00000168769	*TET2*	ENST00000540549
ENSG00000126012	*KDM5C*	ENST00000375401	ENSG00000106799	*TGFBR1*	ENST00000374994
ENSG00000147050	*KDM6A*	ENST00000377967	ENSG00000163513	*TGFBR2*	ENST00000359013
ENSG00000079999	*KEAP1*	ENST00000171111	ENSG00000136869	*TLR4*	ENST00000355622
ENSG00000197993	*KEL*	ENST00000355265	ENSG00000135956	*TMEM127*	ENST00000258439
ENSG00000134313	*KIDINS220*	ENST00000256707	ENSG00000234224	*TMEM229A*	ENST00000455783
ENSG00000157404	*KIT*	ENST00000288135	ENSG00000137747	*TMPRSS13*	ENST00000524993
ENSG00000109787	*KLF3*	ENST00000261438	ENSG00000157873	*TNFRSF14*	ENST00000355716
ENSG00000102554	*KLF5*	ENST00000377687	ENSG00000243509	*TNFRSF6B*	ENST00000369996
ENSG00000205810	*KLRC3*	ENST00000381903	ENSG00000106952	*TNFSF8*	ENST00000223795
ENSG00000167548	*KMT2D*	ENST00000301067	ENSG00000168884	*TNIP2*	ENST00000315423
ENSG00000171798	*KNDC1*	ENST00000304613	ENSG00000141510	*TP53*	ENST00000269305
ENSG00000133703	*KRAS*	ENST00000256078	ENSG00000115705	*TPO*	ENST00000345913
ENSG00000171346	*KRT15*	ENST00000254043	ENSG00000131323	*TRAF3*	ENST00000560371
ENSG00000172867	*KRT2*	ENST00000309680	ENSG00000112195	*TREML2*	ENST00000483722
ENSG00000139648	*KRT71*	ENST00000267119	ENSG00000108395	*TRIM37*	ENST00000262294
ENSG00000161849	*KRT84*	ENST00000257951	ENSG00000147573	*TRIM55*	ENST00000315962
ENSG00000107929	*LARP4B*	ENST00000316157	ENSG00000100106	*TRIOBP*	ENST00000406386
ENSG00000196734	*LCE1B*	ENST00000360090	ENSG00000165699	*TSC1*	ENST00000298552
ENSG00000240386	*LCE1F*	ENST00000334371	ENSG00000103197	*TSC2*	ENST00000219476
ENSG00000187173	*LCE2A*	ENST00000368779	ENSG00000196428	*TSC22D2*	ENST00000361875
ENSG00000163202	*LCE3D*	ENST00000368787	ENSG00000155657	*TTN*	ENST00000589042
ENSG00000169744	*LDB2*	ENST00000304523	ENSG00000077498	*TYR*	ENST00000263321
ENSG00000168924	*LETM1*	ENST00000302787	ENSG00000092445	*TYRO3*	ENST00000263798
ENSG00000050426	*LETMD1*	ENST00000418425	ENSG00000077721	*UBE2A*	ENST00000371558
ENSG00000138039	*LHCGR*	ENST00000294954	ENSG00000083290	*ULK2*	ENST00000395544
ENSG00000239998	*LILRA2*	ENST00000251377	ENSG00000168038	*ULK4*	ENST00000301831
ENSG00000182541	*LIMK2*	ENST00000340552	ENSG00000169021	*UQCRFS1*	ENST00000304863
ENSG00000101670	*LIPG*	ENST00000261292	ENSG00000143258	*USP21*	ENST00000368002
ENSG00000170807	*LMOD2*	ENST00000458573	ENSG00000181408	*UTS2R*	ENST00000313135
ENSG00000203782	*LOR*	ENST00000368742	ENSG00000150630	*VEGFC*	ENST00000280193
ENSG00000144749	*LRIG1*	ENST00000273261	ENSG00000134086	*VHL*	ENST00000256474
ENSG00000131409	*LRRC4B*	ENST00000599957	ENSG00000178201	*VN1R1*	ENST00000321039

ENSG00000148948	*LRRC4C*	ENST00000278198	ENSG00000188730	*VWC2*	ENST00000340652
ENSG00000125872	*LRRN4*	ENST00000378858	ENSG00000015285	*WAS*	ENST00000376701
ENSG00000168056	*LTBP3*	ENST00000301873	ENSG00000239779	*WBP1*	ENST00000233615
ENSG00000062524	*LTK*	ENST00000263800	ENSG00000060237	*WNK1*	ENST00000315939
ENSG00000139329	*LUM*	ENST00000266718	ENSG00000002745	*WNT16*	ENST00000222462
ENSG00000099949	*LZTR1*	ENST00000215739	ENSG00000105989	*WNT2*	ENST00000265441
ENSG00000061337	*LZTS1*	ENST00000381569	ENSG00000165392	*WRN*	ENST00000298139
ENSG00000099866	*MADCAM1*	ENST00000215637	ENSG00000184937	*WT1*	ENST00000332351
ENSG00000130479	*MAP1S*	ENST00000324096	ENSG00000136936	*XPA*	ENST00000375128
ENSG00000065559	*MAP2K4*	ENST00000353533	ENSG00000154767	*XPC*	ENST00000285021
ENSG00000076984	*MAP2K7*	ENST00000397979	ENSG00000181722	*ZBTB20*	ENST00000474710
ENSG00000095015	*MAP3K1*	ENST00000399503	ENSG00000160685	*ZBTB7B*	ENST00000417934
ENSG00000100030	*MAPK1*	ENST00000215832	ENSG00000178199	*ZC3H12D*	ENST00000409806
ENSG00000186868	*MAPT*	ENST00000344290	ENSG00000169554	*ZEB2*	ENST00000558170
ENSG00000125952	*MAX*	ENST00000358664	ENSG00000140836	*ZFHX3*	ENST00000268489
ENSG00000166987	*MBD6*	ENST00000355673	ENSG00000185650	*ZFP36L1*	ENST00000439696
ENSG00000184634	*MED12*	ENST00000374080	ENSG00000152518	*ZFP36L2*	ENST00000282388
ENSG00000133895	*MEN1*	ENST00000337652	ENSG00000179588	*ZFPM1*	ENST00000319555
ENSG00000152595	*MEPE*	ENST00000424957	ENSG00000122515	*ZMIZ2*	ENST00000309315
ENSG00000105976	*MET*	ENST00000318493	ENSG00000105136	*ZNF419*	ENST00000424930
ENSG00000165819	*METTL3*	ENST00000298717	ENSG00000229676	*ZNF492*	ENST00000456783
ENSG00000076242	*MLH1*	ENST00000231790	ENSG00000188171	*ZNF626*	ENST00000601440
ENSG00000171843	*MLLT3*	ENST00000380338	ENSG00000197483	*ZNF628*	ENST00000598519
ENSG00000005381	*MPO*	ENST00000225275	ENSG00000196109	*ZNF676*	ENST00000397121
ENSG00000150054	*MPP7*	ENST00000337532	ENSG00000197123	*ZNF679*	ENST00000421025
ENSG00000095002	*MSH2*	ENST00000233146	ENSG00000182141	*ZNF708*	ENST00000356929
ENSG00000113318	*MSH3*	ENST00000265081	ENSG00000141579	*ZNF750*	ENST00000269394
ENSG00000116062	*MSH6*	ENST00000234420	ENSG00000198146	*ZNF770*	ENST00000356321
ENSG00000163132	*MSX1*	ENST00000382723	ENSG00000204514	*ZNF814*	ENST00000435989
ENSG00000198793	*MTOR*	ENST00000361445	ENSG00000234284	*ZNF879*	ENST00000444149
ENSG00000185499	*MUC1*	ENST00000368395	ENSG00000221923	*ZNF880*	ENST00000422689
ENSG00000169876	*MUC17*	ENST00000306151	ENSG00000188372	*ZP3*	ENST00000394857
ENSG00000132781	*MUTYH*	ENST00000372098	ENSG00000131848	*ZSCAN5A*	ENST00000587340
ENSG00000118513	*MYB*	ENST00000341911	ENSG00000122952	*ZWINT*	ENST00000373944
ENSG00000172936	*MYD88*	ENST00000417037			

Table A6 - Gene list used for analysis of truncating variants based on ratio of non-synonymous variants to synonymous per gene in Martincorena et al. 2017

Gene identifier	Gene name	Canonical transcript	Gene identifier	Gene name	Canonical transcript
ENSG00000135503	*ACVR1B*	ENST00000541224	ENSG00000105663	*KMT2B*	ENST00000607650
ENSG00000121989	*ACVR2A*	ENST00000241416	ENSG00000055609	*KMT2C*	ENST00000262189
ENSG00000129474	*AJUBA*	ENST00000262713	ENSG00000167548	*KMT2D*	ENST00000301067
ENSG00000142208	*AKT1*	ENST00000554581	ENSG00000133703	*KRAS*	ENST00000256078
ENSG00000163631	*ALB*	ENST00000295897	ENSG00000186081	*KRT5*	ENST00000252242
ENSG00000110497	*AMBRA1*	ENST00000314845	ENSG00000150457	*LATS2*	ENST00000382592
ENSG00000184675	*AMER1*	ENST00000330258	ENSG00000099949	*LZTR1*	ENST00000215739
ENSG00000134982	*APC*	ENST00000457016	ENSG00000169032	*MAP2K1*	ENST00000307102
ENSG00000160007	*ARHGAP35*	ENST00000404338	ENSG00000065559	*MAP2K4*	ENST00000353533
ENSG00000100852	*ARHGAP5*	ENST00000345122	ENSG00000076984	*MAP2K7*	ENST00000397979
ENSG00000117713	*ARID1A*	ENST00000324856	ENSG00000095015	*MAP3K1*	ENST00000399503
ENSG00000049618	*ARID1B*	ENST00000346085	ENSG00000125952	*MAX*	ENST00000358664
ENSG00000189079	*ARID2*	ENST00000334344	ENSG00000103495	*MAZ*	ENST00000219782
ENSG00000150347	*ARID5B*	ENST00000279873	ENSG00000166987	*MBD6*	ENST00000355673
ENSG00000171456	*ASXL1*	ENST00000375687	ENSG00000133895	*MEN1*	ENST00000337652
ENSG00000143970	*ASXL2*	ENST00000435504	ENSG00000105976	*MET*	ENST00000318493
ENSG00000149311	*ATM*	ENST00000278616	ENSG00000174197	*MGA*	ENST00000219905
ENSG00000143153	*ATP1B1*	ENST00000367816	ENSG00000076242	*MLH1*	ENST00000231790
ENSG00000085224	*ATRX*	ENST00000373344	ENSG00000095002	*MSH2*	ENST00000233146
ENSG00000103126	*AXIN1*	ENST00000262320	ENSG00000198793	*MTOR*	ENST00000361445
ENSG00000166710	*B2M*	ENST00000558401	ENSG00000184956	*MUC6*	ENST00000421673
ENSG00000163930	*BAP1*	ENST00000460680	ENSG00000136997	*MYC*	ENST00000377970
ENSG00000183337	*BCOR*	ENST00000378444	ENSG00000141027	*NCOR1*	ENST00000268712
ENSG00000204217	*BMPR2*	ENST00000374580	ENSG00000196712	*NF1*	ENST00000358273
ENSG00000157764	*BRAF*	ENST00000288602	ENSG00000186575	*NF2*	ENST00000338641
ENSG00000012048	*BRCA1*	ENST00000471181	ENSG00000116044	*NFE2L2*	ENST00000397062
ENSG00000166164	*BRD7*	ENST00000394689	ENSG00000050344	*NFE2L3*	ENST00000056233
ENSG00000187068	*C3orf70*	ENST00000335012	ENSG00000164190	*NIPBL*	ENST00000282516
ENSG00000064012	*CASP8*	ENST00000358485	ENSG00000148400	*NOTCH1*	ENST00000277541
ENSG00000067955	*CBFB*	ENST00000412916	ENSG00000181163	*NPM1*	ENST00000296930
ENSG00000110092	*CCND1*	ENST00000227507	ENSG00000213281	*NRAS*	ENST00000369535
ENSG00000116815	*CD58*	ENST00000369489	ENSG00000165671	*NSD1*	ENST00000439151
ENSG00000039068	*CDH1*	ENST00000261769	ENSG00000130538	*OR11H1*	ENST00000252835
ENSG00000040731	*CDH10*	ENST00000264463	ENSG00000257115	*OR11H12*	ENST00000550708
ENSG00000167258	*CDK12*	ENST00000447079	ENSG00000176294	*OR4N2*	ENST00000315947
ENSG00000124762	*CDKN1A*	ENST00000405375	ENSG00000163939	*PBRM1*	ENST00000394830
ENSG00000111276	*CDKN1B*	ENST00000228872	ENSG00000101327	*PDYN*	ENST00000217305
ENSG00000147889	*CDKN2A*	ENST00000498124	ENSG00000156531	*PHF6*	ENST00000332070
ENSG00000123080	*CDKN2C*	ENST00000262662	ENSG00000121879	*PIK3CA*	ENST00000263967
ENSG00000245848	*CEBPA*	ENST00000498907	ENSG00000145675	*PIK3R1*	ENST00000521381
ENSG00000111642	*CHD4*	ENST00000357008	ENSG00000221900	*POM121L12*	ENST00000408890
ENSG00000141977	*CIB3*	ENST00000269878	ENSG00000188219	*POTEE*	ENST00000356920
ENSG00000079432	*CIC*	ENST00000575354	ENSG00000170836	*PPM1D*	ENST00000305921
ENSG00000180917	*CMTR2*	ENST00000338099	ENSG00000105568	*PPP2R1A*	ENST00000322088
ENSG00000005339	*CREBBP*	ENST00000262367	ENSG00000138814	*PPP3CA*	ENST00000394854
ENSG00000009307	*CSDE1*	ENST00000438362	ENSG00000119414	*PPP6C*	ENST00000451402
ENSG00000102974	*CTCF*	ENST00000264010	ENSG00000108946	*PRKAR1A*	ENST00000589228
ENSG00000168036	*CTNNB1*	ENST00000349496	ENSG00000167371	*PRRT2*	ENST00000567659

ENSG00000036257	*CUL3*	ENST00000264414	ENSG00000243137	*PSG4*	ENST00000405312
ENSG00000158290	*CUL4B*	ENST00000404115	ENSG00000164985	*PSIP1*	ENST00000380733
ENSG00000257923	*CUX1*	ENST00000360264	ENSG00000171862	*PTEN*	ENST00000371953
ENSG00000160882	*CYP11B1*	ENST00000292427	ENSG00000179295	*PTPN11*	ENST00000351677
ENSG00000071626	*DAZAP1*	ENST00000233078	ENSG00000136238	*RAC1*	ENST00000356142
ENSG00000215301	*DDX3X*	ENST00000399959	ENSG00000161800	*RACGAP1*	ENST00000434422
ENSG00000119772	*DNMT3A*	ENST00000264709	ENSG00000172819	*RARG*	ENST00000425354
ENSG00000010219	*DYRK4*	ENST00000540757	ENSG00000145715	*RASA1*	ENST00000274376
ENSG00000156508	*EEF1A1*	ENST00000316292	ENSG00000139687	*RB1*	ENST00000267163
ENSG00000108947	*EFNB3*	ENST00000226091	ENSG00000182872	*RBM10*	ENST00000377604
ENSG00000146648	*EGFR*	ENST00000275493	ENSG00000067560	*RHOA*	ENST00000418115
ENSG00000173674	*EIF1AX*	ENST00000379607	ENSG00000143878	*RHOB*	ENST00000272233
ENSG00000163435	*ELF3*	ENST00000359651	ENSG00000108375	*RNF43*	ENST00000584437
ENSG00000100393	*EP300*	ENST00000263253	ENSG00000116251	*RPL22*	ENST00000234875
ENSG00000142627	*EPHA2*	ENST00000358432	ENSG00000122406	*RPL5*	ENST00000370321
ENSG00000151491	*EPS8*	ENST00000281172	ENSG00000177189	*RPS6KA3*	ENST00000379565
ENSG00000141736	*ERBB2*	ENST00000269571	ENSG00000124782	*RREB1*	ENST00000379938
ENSG00000065361	*ERBB3*	ENST00000267101	ENSG00000159216	*RUNX1*	ENST00000300305
ENSG00000178568	*ERBB4*	ENST00000342788	ENSG00000186350	*RXRA*	ENST00000481739
ENSG00000104884	*ERCC2*	ENST00000391945	ENSG00000181555	*SETD2*	ENST00000409792
ENSG00000106462	*EZH2*	ENST00000320356	ENSG00000175387	*SMAD2*	ENST00000402690
ENSG00000133193	*FAM104A*	ENST00000405159	ENSG00000141646	*SMAD4*	ENST00000342988
ENSG00000147382	*FAM58A*	ENST00000406277	ENSG00000127616	*SMARCA4*	ENST00000429416
ENSG00000083857	*FAT1*	ENST00000441802	ENSG00000072501	*SMC1A*	ENST00000322213
ENSG00000109670	*FBXW7*	ENST00000281708	ENSG00000108055	*SMC3*	ENST00000361804
ENSG00000066468	*FGFR2*	ENST00000457416	ENSG00000188176	*SMTNL2*	ENST00000389313
ENSG00000068078	*FGFR3*	ENST00000340107	ENSG00000125398	*SOX9*	ENST00000245479
ENSG00000122025	*FLT3*	ENST00000241453	ENSG00000065526	*SPEN*	ENST00000375759
ENSG00000075426	*FOSL2*	ENST00000264716	ENSG00000121067	*SPOP*	ENST00000393331
ENSG00000129514	*FOXA1*	ENST00000250448	ENSG00000197694	*SPTAN1*	ENST00000372739
ENSG00000125798	*FOXA2*	ENST00000419308	ENSG00000101972	*STAG2*	ENST00000218089
ENSG00000114861	*FOXP1*	ENST00000491238	ENSG00000118046	*STK11*	ENST00000326873
ENSG00000164379	*FOXQ1*	ENST00000296839	ENSG00000250264	*TAP2*	ENST00000452392
ENSG00000162613	*FUBP1*	ENST00000370768	ENSG00000135111	*TBX3*	ENST00000257566
ENSG00000224659	*GAGE12J*	ENST00000442437	ENSG00000140262	*TCF12*	ENST00000438423
ENSG00000107485	*GATA3*	ENST00000379328	ENSG00000148737	*TCF7L2*	ENST00000543371
ENSG00000127588	*GNG13*	ENST00000248150	ENSG00000168769	*TET2*	ENST00000540549
ENSG00000132522	*GPS2*	ENST00000380728	ENSG00000042832	*TG*	ENST00000220616
ENSG00000077809	*GTF2I*	ENST00000324896	ENSG00000163513	*TGFBR2*	ENST00000359013
ENSG00000223609	*HBD*	ENST00000380299	ENSG00000177426	*TGIF1*	ENST00000330513
ENSG00000187837	*HIST1H1C*	ENST00000343677	ENSG00000131747	*TOP2A*	ENST00000423485
ENSG00000180573	*HIST1H2AC*	ENST00000602637	ENSG00000141510	*TP53*	ENST00000269305
ENSG00000158373	*HIST1H2BD*	ENST00000289316	ENSG00000165699	*TSC1*	ENST00000298552
ENSG00000124693	*HIST1H3B*	ENST00000244661	ENSG00000103197	*TSC2*	ENST00000219476
ENSG00000184678	*HIST2H2BE*	ENST00000369155	ENSG00000092445	*TYRO3*	ENST00000263798
ENSG00000183598	*HIST2H3D*	ENST00000331491	ENSG00000160201	*U2AF1*	ENST00000291552
ENSG00000206503	*HLA-A*	ENST00000396634	ENSG00000169062	*UPF3A*	ENST00000375299
ENSG00000234745	*HLA-B*	ENST00000412585	ENSG00000048028	*USP28*	ENST00000003302
ENSG00000204525	*HLA-C*	ENST00000376228	ENSG00000134086	*VHL*	ENST00000256474
ENSG00000120093	*HOXB3*	ENST00000470495	ENSG00000184937	*WT1*	ENST00000332351
ENSG00000174775	*HRAS*	ENST00000451590	ENSG00000119596	*YLPM1*	ENST00000325680
ENSG00000138413	*IDH1*	ENST00000415913	ENSG00000181722	*ZBTB20*	ENST00000474710
ENSG00000182054	*IDH2*	ENST00000330062	ENSG00000160685	*ZBTB7B*	ENST00000417934

ENSG00000134352	*IL6ST*	ENST00000381298	ENSG00000140836	*ZFHX3*	ENST00000268489
ENSG00000165458	*INPPL1*	ENST00000298229	ENSG00000185650	*ZFP36L1*	ENST00000439696
ENSG00000162434	*JAK1*	ENST00000342505	ENSG00000152518	*ZFP36L2*	ENST00000282388
ENSG00000120071	*KANSL1*	ENST00000262419	ENSG00000005889	*ZFX*	ENST00000379177
ENSG00000126012	*KDM5C*	ENST00000375401	ENSG00000147130	*ZMYM3*	ENST00000353904
ENSG00000147050	*KDM6A*	ENST00000377967	ENSG00000197123	*ZNF679*	ENST00000421025
ENSG00000079999	*KEAP1*	ENST00000171111	ENSG00000141579	*ZNF750*	ENST00000269394
ENSG00000157404	*KIT*	ENST00000288135	ENSG00000048405	*ZNF800*	ENST00000393313
ENSG00000102554	*KLF5*	ENST00000377687	ENSG00000183579	*ZNRF3*	ENST00000544604
ENSG00000118058	*KMT2A*	ENST00000534358			

Table A7 - Gene list used for analysis of truncating variants based on Gene Ontology terms indicating role in DNA repair

Gene identifier	Gene name	Canonical transcript	Gene identifier	Gene name	Canonical transcript
ENSG00000097007	*ABL1*	ENST00000372348	ENSG00000041880	*PARP3*	ENST00000398755
ENSG00000136518	*ACTL6A*	ENST00000429709	ENSG00000102699	*PARP4*	ENST00000381989
ENSG00000101442	*ACTR5*	ENST00000243903	ENSG00000138496	*PARP9*	ENST00000360356
ENSG00000113812	*ACTR8*	ENST00000335754	ENSG00000185480	*PARPBP*	ENST00000358383
ENSG00000100601	*ALKBH1*	ENST00000216489	ENSG00000157212	*PAXIP1*	ENST00000404141
ENSG00000189046	*ALKBH2*	ENST00000429722	ENSG00000132646	*PCNA*	ENST00000379160
ENSG00000166199	*ALKBH3*	ENST00000302708	ENSG00000127980	*PEX1*	ENST00000248633
ENSG00000125843	*AP5S1*	ENST00000246041	ENSG00000243251	*PGBD3*	ENST00000374127
ENSG00000242802	*AP5Z1*	ENST00000348624	ENSG00000140451	*PIF1*	ENST00000268043
ENSG00000166313	*APBB1*	ENST00000299402	ENSG00000140464	*PML*	ENST00000268058
ENSG00000100823	*APEX1*	ENST00000216714	ENSG00000064933	*PMS1*	ENST00000441310
ENSG00000169188	*APEX2*	ENST00000374987	ENSG00000122512	*PMS2*	ENST00000265849
ENSG00000175279	*APITD1*	ENST00000602787	ENSG00000039650	*PNKP*	ENST00000322344
ENSG00000169621	*APLF*	ENST00000303795	ENSG00000101868	*POLA1*	ENST00000379059
ENSG00000137074	*APTX*	ENST00000379813	ENSG00000070501	*POLB*	ENST00000265421
ENSG00000112249	*ASCC3*	ENST00000369162	ENSG00000062822	*POLD1*	ENST00000440232
ENSG00000111875	*ASF1A*	ENST00000229595	ENSG00000106628	*POLD2*	ENST00000406581
ENSG00000034533	*ASTE1*	ENST00000264992	ENSG00000077514	*POLD3*	ENST00000263681
ENSG00000138138	*ATAD1*	ENST00000308448	ENSG00000175482	*POLD4*	ENST00000312419
ENSG00000215915	*ATAD3C*	ENST00000378785	ENSG00000177084	*POLE*	ENST00000320574
ENSG00000149311	*ATM*	ENST00000278616	ENSG00000100479	*POLE2*	ENST00000216367
ENSG00000175054	*ATR*	ENST00000350721	ENSG00000140521	*POLG*	ENST00000268124
ENSG00000164053	*ATRIP*	ENST00000320211	ENSG00000256525	*POLG2*	ENST00000539111
ENSG00000085224	*ATRX*	ENST00000373344	ENSG00000170734	*POLH*	ENST00000372236
ENSG00000066427	*ATXN3*	ENST00000393287	ENSG00000101751	*POLI*	ENST00000579534
ENSG00000168646	*AXIN2*	ENST00000307078	ENSG00000122008	*POLK*	ENST00000241436
ENSG00000105393	*BABAM1*	ENST00000359435	ENSG00000166169	*POLL*	ENST00000370162
ENSG00000138376	*BARD1*	ENST00000260947	ENSG00000122678	*POLM*	ENST00000242248
ENSG00000009954	*BAZ1B*	ENST00000339594	ENSG00000130997	*POLN*	ENST00000511885
ENSG00000107949	*BCCIP*	ENST00000368759	ENSG00000051341	*POLQ*	ENST00000264233
ENSG00000270181	*BIVM-ERCC5*	ENST00000602836	ENSG00000181222	*POLR2A*	ENST00000322644
ENSG00000197299	*BLM*	ENST00000355112	ENSG00000047315	*POLR2B*	ENST00000381227
ENSG00000012048	*BRCA1*	ENST00000471181	ENSG00000102978	*POLR2C*	ENST00000219252
ENSG00000139618	*BRCA2*	ENST00000544455	ENSG00000144231	*POLR2D*	ENST00000272645
ENSG00000185515	*BRCC3*	ENST00000369462	ENSG00000099817	*POLR2E*	ENST00000215587
ENSG00000158019	*BRE*	ENST00000344773	ENSG00000100142	*POLR2F*	ENST00000442738
ENSG00000136492	*BRIP1*	ENST00000259008	ENSG00000168002	*POLR2G*	ENST00000301788
ENSG00000159388	*BTG2*	ENST00000290551	ENSG00000163882	*POLR2H*	ENST00000456318
ENSG00000158636	*C11orf30*	ENST00000529032	ENSG00000105258	*POLR2I*	ENST00000221859
ENSG00000185504	*C17orf70*	ENST00000327787	ENSG00000005075	*POLR2J*	ENST00000292614
ENSG00000131944	*C19orf40*	ENST00000588258	ENSG00000147669	*POLR2K*	ENST00000353107
ENSG00000162585	*C1orf86*	ENST00000378546	ENSG00000177700	*POLR2L*	ENST00000322028
ENSG00000134480	*CCNH*	ENST00000256897	ENSG00000149923	*PPP4C*	ENST00000279387
ENSG00000152669	*CCNO*	ENST00000282572	ENSG00000163605	*PPP4R2*	ENST00000356692
ENSG00000081377	*CDC14B*	ENST00000375241	ENSG00000011485	*PPP5C*	ENST00000012443
ENSG00000146670	*CDCA5*	ENST00000275517	ENSG00000164306	*PRIMPOL*	ENST00000314970
ENSG00000170312	*CDK1*	ENST00000395284	ENSG00000126583	*PRKCG*	ENST00000263431
ENSG00000123374	*CDK2*	ENST00000266970	ENSG00000253729	*PRKDC*	ENST00000314191
ENSG00000134058	*CDK7*	ENST00000256443	ENSG00000198890	*PRMT6*	ENST00000370078
ENSG00000136807	*CDK9*	ENST00000373264	ENSG00000110107	*PRPF19*	ENST00000227524

ENSG00000129355	*CDKN2D*	ENST00000393599	ENSG00000100764	*PSMC1*	ENST00000261303
ENSG00000153879	*CEBPG*	ENST00000284000	ENSG00000165916	*PSMC3*	ENST00000298852
ENSG00000110274	*CEP164*	ENST00000278935	ENSG00000013275	*PSMC4*	ENST00000157812
ENSG00000147400	*CETN2*	ENST00000370277	ENSG00000087191	*PSMC5*	ENST00000310144
ENSG00000167670	*CHAF1A*	ENST00000301280	ENSG00000100519	*PSMC6*	ENST00000445930
ENSG00000159259	*CHAF1B*	ENST00000314103	ENSG00000115233	*PSMD14*	ENST00000409682
ENSG00000131778	*CHD1L*	ENST00000369258	ENSG00000068878	*PSME4*	ENST00000404125
ENSG00000149554	*CHEK1*	ENST00000534070	ENSG00000164611	*PTTG1*	ENST00000393964
ENSG00000183765	*CHEK2*	ENST00000382580	ENSG00000113456	*RAD1*	ENST00000382038
ENSG00000101204	*CHRNA4*	ENST00000370263	ENSG00000152942	*RAD17*	ENST00000509734
ENSG00000127586	*CHTF18*	ENST00000262315	ENSG00000070950	*RAD18*	ENST00000264926
ENSG00000185043	*CIB1*	ENST00000328649	ENSG00000164754	*RAD21*	ENST00000297338
ENSG00000100865	*CINP*	ENST00000536961	ENSG00000244588	*RAD21L1*	ENST00000409241
ENSG00000092853	*CLSPN*	ENST00000318121	ENSG00000179262	*RAD23A*	ENST00000586534
ENSG00000008405	*CRY1*	ENST00000008527	ENSG00000119318	*RAD23B*	ENST00000358015
ENSG00000121671	*CRY2*	ENST00000443527	ENSG00000113522	*RAD50*	ENST00000265335
ENSG00000141551	*CSNK1D*	ENST00000314028	ENSG00000051180	*RAD51*	ENST00000382643
ENSG00000213923	*CSNK1E*	ENST00000396832	ENSG00000111247	*RAD51AP1*	ENST00000228843
ENSG00000269307	*CTD-2278I10.6*	ENST00000596542	ENSG00000182185	*RAD51B*	ENST00000487270
ENSG00000139842	*CUL4A*	ENST00000375440	ENSG00000108384	*RAD51C*	ENST00000337432
ENSG00000158290	*CUL4B*	ENST00000404115	ENSG00000185379	*RAD51D*	ENST00000590016
ENSG00000198924	*DCLRE1A*	ENST00000361384	ENSG00000002016	*RAD52*	ENST00000358495
ENSG00000118655	*DCLRE1B*	ENST00000369563	ENSG00000197275	*RAD54B*	ENST00000336148
ENSG00000152457	*DCLRE1C*	ENST00000378278	ENSG00000085999	*RAD54L*	ENST00000371975
ENSG00000167986	*DDB1*	ENST00000301764	ENSG00000172613	*RAD9A*	ENST00000307980
ENSG00000134574	*DDB2*	ENST00000256996	ENSG00000151164	*RAD9B*	ENST00000392672
ENSG00000079785	*DDX1*	ENST00000381341	ENSG00000101773	*RBBP8*	ENST00000399722
ENSG00000124795	*DEK*	ENST00000397239	ENSG00000239306	*RBM14*	ENST00000310137
ENSG00000178028	*DMAP1*	ENST00000372289	ENSG00000100387	*RBX1*	ENST00000216225
ENSG00000100206	*DMC1*	ENST00000216024	ENSG00000187456	*RDM1*	ENST00000293273
ENSG00000138346	*DNA2*	ENST00000399180	ENSG00000100918	*REC8*	ENST00000311457
ENSG00000143476	*DTL*	ENST00000366991	ENSG00000004700	*RECQL*	ENST00000444129
ENSG00000163840	*DTX3L*	ENST00000296161	ENSG00000108469	*RECQL5*	ENST00000317905
ENSG00000122547	*EEPD1*	ENST00000242108	ENSG00000135945	*REV1*	ENST00000258428
ENSG00000146648	*EGFR*	ENST00000275493	ENSG00000009413	*REV3L*	ENST00000358835
ENSG00000255150	*EID3*	ENST00000527879	ENSG00000035928	*RFC1*	ENST00000381897
ENSG00000154920	*EME1*	ENST00000393271	ENSG00000049541	*RFC2*	ENST00000055077
ENSG00000197774	*EME2*	ENST00000568449	ENSG00000133119	*RFC3*	ENST00000380071
ENSG00000173818	*ENDOV*	ENST00000518137	ENSG00000163918	*RFC4*	ENST00000392481
ENSG00000135999	*EPC2*	ENST00000258484	ENSG00000111445	*RFC5*	ENST00000454402
ENSG00000012061	*ERCC1*	ENST00000013807	ENSG00000168411	*RFWD3*	ENST00000361070
ENSG00000104884	*ERCC2*	ENST00000391945	ENSG00000171792	*RHNO1*	ENST00000489288
ENSG00000163161	*ERCC3*	ENST00000285398	ENSG00000163961	*RNF168*	ENST00000318037
ENSG00000175595	*ERCC4*	ENST00000311895	ENSG00000166439	*RNF169*	ENST00000299563
ENSG00000134899	*ERCC5*	ENST00000355739	ENSG00000112130	*RNF8*	ENST00000373479
ENSG00000225830	*ERCC6*	ENST00000355832	ENSG00000260914	*RP11-343C2.11*	ENST00000570054
ENSG00000182150	*ERCC6L2*	ENST00000288985	ENSG00000254469	*RP11-849H4.2*	ENST00000528511
ENSG00000258838	*ERCC6-PGBD3*	ENST00000515869	ENSG00000132383	*RPA1*	ENST00000254719
ENSG00000049167	*ERCC8*	ENST00000265038	ENSG00000117748	*RPA2*	ENST00000373912
ENSG00000174371	*EXO1*	ENST00000366548	ENSG00000106399	*RPA3*	ENST00000223129
ENSG00000164002	*EXO5*	ENST00000372703	ENSG00000204086	*RPA4*	ENST00000373040
ENSG00000104313	*EYA1*	ENST00000340726	ENSG00000129197	*RPAIN*	ENST00000405578

ENSG00000064655	*EYA2*	ENST00000327619	ENSG00000143947	*RPS27A*	ENST00000272317
ENSG00000158161	*EYA3*	ENST00000373871	ENSG00000185088	*RPS27L*	ENST00000330964
ENSG00000112319	*EYA4*	ENST00000367895	ENSG00000149273	*RPS3*	ENST00000278572
ENSG00000163322	*FAM175A*	ENST00000321945	ENSG00000048392	*RRM2B*	ENST00000251810
ENSG00000198690	*FAN1*	ENST00000362065	ENSG00000258366	*RTEL1*	ENST00000508582
ENSG00000187741	*FANCA*	ENST00000389301	ENSG00000175792	*RUVBL1*	ENST00000322623
ENSG00000181544	*FANCB*	ENST00000398334	ENSG00000183207	*RUVBL2*	ENST00000595090
ENSG00000158169	*FANCC*	ENST00000289081	ENSG00000181555	*SETD2*	ENST00000409792
ENSG00000144554	*FANCD2*	ENST00000287647	ENSG00000170364	*SETMAR*	ENST00000358065
ENSG00000112039	*FANCE*	ENST00000229769	ENSG00000107290	*SETX*	ENST00000224140
ENSG00000183161	*FANCF*	ENST00000327470	ENSG00000116560	*SFPQ*	ENST00000357214
ENSG00000221829	*FANCG*	ENST00000378643	ENSG00000156384	*SFR1*	ENST00000369727
ENSG00000140525	*FANCI*	ENST00000310775	ENSG00000127922	*SHFM1*	ENST00000248566
ENSG00000115392	*FANCL*	ENST00000402135	ENSG00000146414	*SHPRH*	ENST00000367505
ENSG00000187790	*FANCM*	ENST00000267430	ENSG00000096717	*SIRT1*	ENST00000212015
ENSG00000116663	*FBXO6*	ENST00000376753	ENSG00000077463	*SIRT6*	ENST00000337491
ENSG00000168496	*FEN1*	ENST00000305885	ENSG00000014824	*SLC30A9*	ENST00000264451
ENSG00000070193	*FGF10*	ENST00000264664	ENSG00000132207	*SLX1A*	ENST00000251303
ENSG00000132436	*FIGNL1*	ENST00000419119	ENSG00000181625	*SLX1B*	ENST00000330181
ENSG00000111206	*FOXM1*	ENST00000342628	ENSG00000188827	*SLX4*	ENST00000294008
ENSG00000140718	*FTO*	ENST00000471389	ENSG00000153147	*SMARCA5*	ENST00000283131
ENSG00000105325	*FZR1*	ENST00000395095	ENSG00000163104	*SMARCAD1*	ENST00000359052
ENSG00000116717	*GADD45A*	ENST00000370986	ENSG00000099956	*SMARCB1*	ENST00000263121
ENSG00000178295	*GEN1*	ENST00000381254	ENSG00000072501	*SMC1A*	ENST00000322213
ENSG00000110768	*GTF2H1*	ENST00000265963	ENSG00000077935	*SMC1B*	ENST00000357450
ENSG00000145736	*GTF2H2*	ENST00000330280	ENSG00000136824	*SMC2*	ENST00000286398
ENSG00000183474	*GTF2H2C*	ENST00000510979	ENSG00000108055	*SMC3*	ENST00000361804
ENSG00000262261	*GTF2H2D*	ENST00000577126	ENSG00000113810	*SMC4*	ENST00000357388
ENSG00000111358	*GTF2H3*	ENST00000543341	ENSG00000198887	*SMC5*	ENST00000361138
ENSG00000213780	*GTF2H4*	ENST00000259895	ENSG00000163029	*SMC6*	ENST00000448223
ENSG00000272047	*GTF2H5*	ENST00000607778	ENSG00000157106	*SMG1*	ENST00000446231
ENSG00000188486	*H2AFX*	ENST00000530167	ENSG00000123415	*SMUG1*	ENST00000508394
ENSG00000128731	*HERC2*	ENST00000261609	ENSG00000142168	*SOD1*	ENST00000270142
ENSG00000172273	*HINFP*	ENST00000350777	ENSG00000021574	*SPAST*	ENST00000315285
ENSG00000137309	*HMGA1*	ENST00000447654	ENSG00000141255	*SPATA22*	ENST00000573128
ENSG00000149948	*HMGA2*	ENST00000403681	ENSG00000145375	*SPATA5*	ENST00000274008
ENSG00000189403	*HMGB1*	ENST00000405805	ENSG00000171763	*SPATA5L1*	ENST00000305560
ENSG00000164104	*HMGB2*	ENST00000296503	ENSG00000164808	*SPIDR*	ENST00000297423
ENSG00000205581	*HMGN1*	ENST00000380749	ENSG00000010072	*SPRTN*	ENST00000295050
ENSG00000136273	*HUS1*	ENST00000258774	ENSG00000149136	*SSRP1*	ENST00000278412
ENSG00000188996	*HUS1B*	ENST00000380907	ENSG00000169689	*STRA13*	ENST00000392359
ENSG00000086758	*HUWE1*	ENST00000342160	ENSG00000103266	*STUB1*	ENST00000219548
ENSG00000137331	*IER3*	ENST00000259874	ENSG00000116030	*SUMO1*	ENST00000392246
ENSG00000132740	*IGHMBP2*	ENST00000255078	ENSG00000092201	*SUPT16H*	ENST00000216297
ENSG00000148153	*INIP*	ENST00000374242	ENSG00000175854	*SWI5*	ENST00000320188
ENSG00000128908	*INO80*	ENST00000361937	ENSG00000173928	*SWSAP1*	ENST00000312423
ENSG00000115274	*INO80B*	ENST00000233331	ENSG00000160551	*TAOK1*	ENST00000261716
ENSG00000153391	*INO80C*	ENST00000441607	ENSG00000135090	*TAOK3*	ENST00000392533
ENSG00000114933	*INO80D*	ENST00000403263	ENSG00000187735	*TCEA1*	ENST00000521604
ENSG00000169592	*INO80E*	ENST00000563197	ENSG00000139372	*TDG*	ENST00000392872
ENSG00000143624	*INTS3*	ENST00000318967	ENSG00000042088	*TDP1*	ENST00000335725
ENSG00000152409	*JMY*	ENST00000396137	ENSG00000111802	*TDP2*	ENST00000378198
ENSG00000172977	*KAT5*	ENST00000341318	ENSG00000166848	*TERF2IP*	ENST00000300086
ENSG00000186625	*KATNA1*	ENST00000367411	ENSG00000133863	*TEX15*	ENST00000256246
ENSG00000102781	*KATNAL1*	ENST00000380615	ENSG00000105619	*TFPT*	ENST00000391759
ENSG00000167216	*KATNAL2*	ENST00000245121	ENSG00000140534	*TICRR*	ENST00000268138

ENSG00000166803	*KIAA0101*	ENST00000300035	ENSG00000064545	*TMEM161A*	ENST00000162044
ENSG00000166783	*KIAA0430*	ENST00000396368	ENSG00000118245	*TNP1*	ENST00000236979
ENSG00000050030	*KIAA2022*	ENST00000055682	ENSG00000260716	*TONSL*	ENST00000409379
ENSG00000079616	*KIF22*	ENST00000160827	ENSG00000131747	*TOP2A*	ENST00000423485
ENSG00000151657	*KIN*	ENST00000379562	ENSG00000163781	*TOPBP1*	ENST00000260810
ENSG00000268361	*L34079.2*	ENST00000594374	ENSG00000141510	*TP53*	ENST00000269305
ENSG00000105486	*LIG1*	ENST00000263274	ENSG00000067369	*TP53BP1*	ENST00000382044
ENSG00000005156	*LIG3*	ENST00000378526	ENSG00000078900	*TP73*	ENST00000378295
ENSG00000174405	*LIG4*	ENST00000356922	ENSG00000213689	*TREX1*	ENST00000422277
ENSG00000196365	*LONP1*	ENST00000360614	ENSG00000183479	*TREX2*	ENST00000330912
ENSG00000116670	*MAD2L2*	ENST00000235310	ENSG00000130726	*TRIM28*	ENST00000253024
ENSG00000129071	*MBD4*	ENST00000249910	ENSG00000153827	*TRIP12*	ENST00000283943
ENSG00000258839	*MC1R*	ENST00000555147	ENSG00000071539	*TRIP13*	ENST00000166345
ENSG00000125885	*MCM8*	ENST00000378896	ENSG00000136319	*TTC5*	ENST00000258821
ENSG00000111877	*MCM9*	ENST00000316316	ENSG00000122691	*TWIST1*	ENST00000242261
ENSG00000187778	*MCRS1*	ENST00000357123	ENSG00000176890	*TYMS*	ENST00000323274
ENSG00000137337	*MDC1*	ENST00000376406	ENSG00000221983	*UBA52*	ENST00000442744
ENSG00000162039	*MEIOB*	ENST00000412554	ENSG00000170315	*UBB*	ENST00000302182
ENSG00000133895	*MEN1*	ENST00000337652	ENSG00000150991	*UBC*	ENST00000536769
ENSG00000125871	*MGME1*	ENST00000377710	ENSG00000077721	*UBE2A*	ENST00000371558
ENSG00000170430	*MGMT*	ENST00000306010	ENSG00000119048	*UBE2B*	ENST00000265339
ENSG00000076242	*MLH1*	ENST00000231790	ENSG00000109332	*UBE2D3*	ENST00000357194
ENSG00000119684	*MLH3*	ENST00000355774	ENSG00000177889	*UBE2N*	ENST00000318066
ENSG00000155229	*MMS19*	ENST00000438925	ENSG00000077152	*UBE2T*	ENST00000367274
ENSG00000146263	*MMS22L*	ENST00000275053	ENSG00000244687	*UBE2V1*	ENST00000340309
ENSG00000020426	*MNAT1*	ENST00000261245	ENSG00000169139	*UBE2V2*	ENST00000523111
ENSG00000185787	*MORF4L1*	ENST00000331268	ENSG00000104343	*UBE2W*	ENST00000419880
ENSG00000123562	*MORF4L2*	ENST00000423833	ENSG00000135018	*UBQLN1*	ENST00000376395
ENSG00000103152	*MPG*	ENST00000219431	ENSG00000188021	*UBQLN2*	ENST00000338222
ENSG00000020922	*MRE11A*	ENST00000323929	ENSG00000160803	*UBQLN4*	ENST00000368309
ENSG00000095002	*MSH2*	ENST00000233146	ENSG00000104517	*UBR5*	ENST00000520539
ENSG00000113318	*MSH3*	ENST00000265081	ENSG00000116750	*UCHL5*	ENST00000367455
ENSG00000057468	*MSH4*	ENST00000263187	ENSG00000087206	*UIMC1*	ENST00000377227
ENSG00000204410	*MSH5*	ENST00000375703	ENSG00000076248	*UNG*	ENST00000242576
ENSG00000255152	*MSH5-SAPCD1*	ENST00000493662	ENSG00000005007	*UPF1*	ENST00000262803
ENSG00000116062	*MSH6*	ENST00000234420	ENSG00000162607	*USP1*	ENST00000339950
ENSG00000160953	*MUM1*	ENST00000344663	ENSG00000103194	*USP10*	ENST00000219473
ENSG00000172732	*MUS81*	ENST00000308110	ENSG00000048028	*USP28*	ENST00000003302
ENSG00000132781	*MUTYH*	ENST00000372098	ENSG00000140455	*USP3*	ENST00000380324
ENSG00000173559	*NABP1*	ENST00000425611	ENSG00000170242	*USP47*	ENST00000339865
ENSG00000139579	*NABP2*	ENST00000380198	ENSG00000187555	*USP7*	ENST00000344836
ENSG00000104320	*NBN*	ENST00000265433	ENSG00000198382	*UVRAG*	ENST00000356136
ENSG00000198646	*NCOA6*	ENST00000374796	ENSG00000163945	*UVSSA*	ENST00000389851
ENSG00000185115	*NDNL2*	ENST00000332303	ENSG00000165280	*VCP*	ENST00000358901
ENSG00000140398	*NEIL1*	ENST00000564784	ENSG00000132612	*VPS4A*	ENST00000254950
ENSG00000154328	*NEIL2*	ENST00000284503	ENSG00000119541	*VPS4B*	ENST00000238497
ENSG00000109674	*NEIL3*	ENST00000264596	ENSG00000136709	*WDR33*	ENST00000322313
ENSG00000170322	*NFRKB*	ENST00000524794	ENSG00000165392	*WRN*	ENST00000298139
ENSG00000151092	*NGLY1*	ENST00000280700	ENSG00000124535	*WRNIP1*	ENST00000380773
ENSG00000187736	*NHEJ1*	ENST00000356853	ENSG00000076924	*XAB2*	ENST00000358368
ENSG00000147140	*NONO*	ENST00000276079	ENSG00000136936	*XPA*	ENST00000375128
ENSG00000181163	*NPM1*	ENST00000296930	ENSG00000154767	*XPC*	ENST00000285021
ENSG00000169189	*NSMCE1*	ENST00000361439	ENSG00000073050	*XRCC1*	ENST00000262887
ENSG00000156831	*NSMCE2*	ENST00000287437	ENSG00000196584	*XRCC2*	ENST00000359321
ENSG00000107672	*NSMCE4A*	ENST00000369023	ENSG00000126215	*XRCC3*	ENST00000553264

ENSG00000065057	*NTHL1*	ENST00000219066	ENSG00000152422	*XRCC4*	ENST00000511817
ENSG00000106268	*NUDT1*	ENST00000397049	ENSG00000079246	*XRCC5*	ENST00000392133
ENSG00000198585	*NUDT16*	ENST00000502852	ENSG00000196419	*XRCC6*	ENST00000359308
ENSG00000143748	*NVL*	ENST00000281701	ENSG00000166896	*XRCC6BP1*	ENST00000300145
ENSG00000114026	*OGG1*	ENST00000302036	ENSG00000100811	*YY1*	ENST00000262238
ENSG00000167770	*OTUB1*	ENST00000538426	ENSG00000011590	*ZBTB32*	ENST00000392197
ENSG00000083093	*PALB2*	ENST00000261584	ENSG00000072121	*ZFYVE26*	ENST00000347230
ENSG00000112941	*PAPD7*	ENST00000230859	ENSG00000121988	*ZRANB3*	ENST00000264159
ENSG00000143799	*PARP1*	ENST00000366794	ENSG00000214941	*ZSWIM7*	ENST00000399277
ENSG00000129484	*PARP2*	ENST00000250416			

Table A8 - Gene list used for analysis of truncating variants based on interactions with known CPGs in GeneMania

Gene identifier	Gene name	Canonical transcript	Gene identifier	Gene name	Canonical transcript
ENSG00000073734	*ABCB11*	ENST00000263817	ENSG00000133895	*MEN1*	ENST00000337652
ENSG00000106546	*AHR*	ENST00000242057	ENSG00000076242	*MLH1*	ENST00000231790
ENSG00000110711	*AIP*	ENST00000279146	ENSG00000119684	*MLH3*	ENST00000355774
ENSG00000134982	*APC*	ENST00000457016	ENSG00000251349	*MSANTD3-TMEFF1*	ENST00000502978
ENSG00000100823	*APEX1*	ENST00000216714	ENSG00000095002	*MSH2*	ENST00000233146
ENSG00000149311	*ATM*	ENST00000278616	ENSG00000116062	*MSH6*	ENST00000234420
ENSG00000168646	*AXIN2*	ENST00000307078	ENSG00000099810	*MTAP*	ENST00000380172
ENSG00000163930	*BAP1*	ENST00000460680	ENSG00000132781	*MUTYH*	ENST00000372098
ENSG00000138376	*BARD1*	ENST00000260947	ENSG00000104320	*NBN*	ENST00000265433
ENSG00000197299	*BLM*	ENST00000355112	ENSG00000196712	*NF1*	ENST00000358273
ENSG00000107779	*BMPR1A*	ENST00000372037	ENSG00000186575	*NF2*	ENST00000338641
ENSG00000012048	*BRCA1*	ENST00000471181	ENSG00000239672	*NME1*	ENST00000336097
ENSG00000139618	*BRCA2*	ENST00000544455	ENSG00000165671	*NSD1*	ENST00000439151
ENSG00000136492	*BRIP1*	ENST00000259008	ENSG00000065057	*NTHL1*	ENST00000219066
ENSG00000156970	*BUB1B*	ENST00000287598	ENSG00000083093	*PALB2*	ENST00000261584
ENSG00000185504	*C17orf70*	ENST00000327787	ENSG00000109132	*PHOX2B*	ENST00000226382
ENSG00000131944	*C19orf40*	ENST00000588258	ENSG00000064933	*PMS1*	ENST00000441310
ENSG00000110395	*CBL*	ENST00000264033	ENSG00000122512	*PMS2*	ENST00000265849
ENSG00000110092	*CCND1*	ENST00000227507	ENSG00000062822	*POLD1*	ENST00000440232
ENSG00000134371	*CDC73*	ENST00000367435	ENSG00000106628	*POLD2*	ENST00000406581
ENSG00000039068	*CDH1*	ENST00000261769	ENSG00000177084	*POLE*	ENST00000320574
ENSG00000105810	*CDK6*	ENST00000265734	ENSG00000170734	*POLH*	ENST00000372236
ENSG00000111276	*CDKN1B*	ENST00000228872	ENSG00000180644	*PRF1*	ENST00000441259
ENSG00000129757	*CDKN1C*	ENST00000414822	ENSG00000108946	*PRKAR1A*	ENST00000589228
ENSG00000147889	*CDKN2A*	ENST00000498124	ENSG00000204983	*PRSS1*	ENST00000311737
ENSG00000147883	*CDKN2B*	ENST00000276925	ENSG00000185920	*PTCH1*	ENST00000331920
ENSG00000245848	*CEBPA*	ENST00000498907	ENSG00000171862	*PTEN*	ENST00000371953
ENSG00000166037	*CEP57*	ENST00000325542	ENSG00000051180	*RAD51*	ENST00000382643
ENSG00000183765	*CHEK2*	ENST00000382580	ENSG00000182185	*RAD51B*	ENST00000487270
ENSG00000114270	*COL7A1*	ENST00000328333	ENSG00000108384	*RAD51C*	ENST00000337432
ENSG00000083799	*CYLD*	ENST00000427738	ENSG00000185379	*RAD51D*	ENST00000590016
ENSG00000167986	*DDB1*	ENST00000301764	ENSG00000139687	*RB1*	ENST00000267163
ENSG00000134574	*DDB2*	ENST00000256996	ENSG00000160957	*RECQL4*	ENST00000428558
ENSG00000100697	*DICER1*	ENST00000526495	ENSG00000132383	*RPA1*	ENST00000254719
ENSG00000144535	*DIS3L2*	ENST00000325385	ENSG00000159216	*RUNX1*	ENST00000300305
ENSG00000130826	*DKC1*	ENST00000369550	ENSG00000126524	*SBDS*	ENST00000246868
ENSG00000107099	*DOCK8*	ENST00000453981	ENSG00000073578	*SDHA*	ENST00000264932
ENSG00000197561	*ELANE*	ENST00000590230	ENSG00000167985	*SDHAF2*	ENST00000301761
ENSG00000119888	*EPCAM*	ENST00000263735	ENSG00000117118	*SDHB*	ENST00000375499
ENSG00000012061	*ERCC1*	ENST00000013807	ENSG00000143252	*SDHC*	ENST00000367975
ENSG00000104884	*ERCC2*	ENST00000391945	ENSG00000204370	*SDHD*	ENST00000375549
ENSG00000163161	*ERCC3*	ENST00000285398	ENSG00000197249	*SERPINA1*	ENST00000448921
ENSG00000175595	*ERCC4*	ENST00000311895	ENSG00000183918	*SH2D1A*	ENST00000371139
ENSG00000134899	*ERCC5*	ENST00000355739	ENSG00000004864	*SLC25A13*	ENST00000416240
ENSG00000182197	*EXT1*	ENST00000378204	ENSG00000188827	*SLX4*	ENST00000294008
ENSG00000151348	*EXT2*	ENST00000395673	ENSG00000141646	*SMAD4*	ENST00000342988
ENSG00000106462	*EZH2*	ENST00000320356	ENSG00000127616	*SMARCA4*	ENST00000429416
ENSG00000103876	*FAH*	ENST00000407106	ENSG00000099956	*SMARCB1*	ENST00000263121
ENSG00000187741	*FANCA*	ENST00000389301	ENSG00000073584	*SMARCE1*	ENST00000348513
ENSG00000181544	*FANCB*	ENST00000398334	ENSG00000184895	*SRY*	ENST00000383070

ENSG00000158169	*FANCC*	ENST00000289081	ENSG00000168610	*STAT3*	ENST00000264657
ENSG00000144554	*FANCD2*	ENST00000287647	ENSG00000118046	*STK11*	ENST00000326873
ENSG00000112039	*FANCE*	ENST00000229769	ENSG00000107882	*SUFU*	ENST00000369902
ENSG00000183161	*FANCF*	ENST00000327470	ENSG00000139546	*TARBP2*	ENST00000266987
ENSG00000221829	*FANCG*	ENST00000378643	ENSG00000132604	*TERF2*	ENST00000603068
ENSG00000140525	*FANCI*	ENST00000310775	ENSG00000164362	*TERT*	ENST00000310581
ENSG00000115392	*FANCL*	ENST00000402135	ENSG00000106799	*TGFBR1*	ENST00000374994
ENSG00000187790	*FANCM*	ENST00000267430	ENSG00000135956	*TMEM127*	ENST00000258439
ENSG00000026103	*FAS*	ENST00000355740	ENSG00000177302	*TOP3A*	ENST00000321105
ENSG00000091483	*FH*	ENST00000366560	ENSG00000141510	*TP53*	ENST00000269305
ENSG00000154803	*FLCN*	ENST00000285071	ENSG00000108395	*TRIM37*	ENST00000262294
ENSG00000179348	*GATA2*	ENST00000341105	ENSG00000165699	*TSC1*	ENST00000298552
ENSG00000177628	*GBA*	ENST00000327247	ENSG00000103197	*TSC2*	ENST00000219476
ENSG00000169562	*GJB1*	ENST00000374022	ENSG00000126088	*UROD*	ENST00000246337
ENSG00000165474	*GJB2*	ENST00000382844	ENSG00000134086	*VHL*	ENST00000256474
ENSG00000147257	*GPC3*	ENST00000394299	ENSG00000015285	*WAS*	ENST00000376701
ENSG00000010704	*HFE*	ENST00000357618	ENSG00000165392	*WRN*	ENST00000298139
ENSG00000256269	*HMBS*	ENST00000278715	ENSG00000184937	*WT1*	ENST00000332351
ENSG00000135100	*HNF1A*	ENST00000257555	ENSG00000136936	*XPA*	ENST00000375128
ENSG00000113263	*ITK*	ENST00000422843	ENSG00000154767	*XPC*	ENST00000285021
ENSG00000125952	*MAX*	ENST00000358664	ENSG00000126215	*XRCC3*	ENST00000553264

Table A9 - Known and possible proto-oncogene cancer predisposition genes used for analysis

Gene identifier	Gene name	Canonical transcript	Gene identifier	Gene name	Canonical transcript
ENSG00000181409	*AATK*	ENST00000326724	ENSG00000198399	*ITSN2*	ENST00000355123
ENSG00000159842	*ABR*	ENST00000302538	ENSG00000160145	*KALRN*	ENST00000240874
ENSG00000170776	*AKAP13*	ENST00000361243	ENSG00000128052	*KDR*	ENST00000263923
ENSG00000171094	*ALK*	ENST00000389048	ENSG00000157404	*KIT*	ENST00000288135
ENSG00000003393	*ALS2*	ENST00000264276	ENSG00000133703	*KRAS*	ENST00000256078
ENSG00000076928	*ARHGEF1*	ENST00000337665	ENSG00000164715	*LMTK2*	ENST00000297293
ENSG00000104728	*ARHGEF10*	ENST00000349830	ENSG00000142235	*LMTK3*	ENST00000270238
ENSG00000074964	*ARHGEF10L*	ENST00000361221	ENSG00000062524	*LTK*	ENST00000263800
ENSG00000132694	*ARHGEF11*	ENST00000368194	ENSG00000101977	*MCF2*	ENST00000519895
ENSG00000196914	*ARHGEF12*	ENST00000397843	ENSG00000126217	*MCF2L*	ENST00000535094
ENSG00000198844	*ARHGEF15*	ENST00000361926	ENSG00000053524	*MCF2L2*	ENST00000328913
ENSG00000130762	*ARHGEF16*	ENST00000378378	ENSG00000153208	*MERTK*	ENST00000295408
ENSG00000110237	*ARHGEF17*	ENST00000263674	ENSG00000105976	*MET*	ENST00000318493
ENSG00000104880	*ARHGEF18*	ENST00000359920	ENSG00000158186	*MRAS*	ENST00000289104
ENSG00000142632	*ARHGEF19*	ENST00000270747	ENSG00000164078	*MST1R*	ENST00000296474
ENSG00000116584	*ARHGEF2*	ENST00000361247	ENSG00000030304	*MUSK*	ENST00000374448
ENSG00000240771	*ARHGEF25*	ENST00000333972	ENSG00000173848	*NET1*	ENST00000355029
ENSG00000114790	*ARHGEF26*	ENST00000356448	ENSG00000066248	*NGEF*	ENST00000264051
ENSG00000163947	*ARHGEF3*	ENST00000338458	ENSG00000197885	*NKIRAS1*	ENST00000443659
ENSG00000214694	*ARHGEF33*	ENST00000409978	ENSG00000168256	*NKIRAS2*	ENST00000307641
ENSG00000204959	*ARHGEF34P*	ENST00000378112	ENSG00000213281	*NRAS*	ENST00000369535
ENSG00000213214	*ARHGEF35*	ENST00000378115	ENSG00000198400	*NTRK1*	ENST00000524377
ENSG00000183111	*ARHGEF37*	ENST00000333677	ENSG00000148053	*NTRK2*	ENST00000376214
ENSG00000236699	*ARHGEF38*	ENST00000420470	ENSG00000140538	*NTRK3*	ENST00000360948
ENSG00000136002	*ARHGEF4*	ENST00000326016	ENSG00000154358	*OBSCN*	ENST00000570156
ENSG00000165801	*ARHGEF40*	ENST00000298694	ENSG00000134853	*PDGFRA*	ENST00000257290
ENSG00000050327	*ARHGEF5*	ENST00000056217	ENSG00000113721	*PDGFRB*	ENST00000261799
ENSG00000129675	*ARHGEF6*	ENST00000250617	ENSG00000120278	*PLEKHG1*	ENST00000367328
ENSG00000102606	*ARHGEF7*	ENST00000375741	ENSG00000090924	*PLEKHG2*	ENST00000409794
ENSG00000131089	*ARHGEF9*	ENST00000253401	ENSG00000126822	*PLEKHG3*	ENST00000247226
ENSG00000167601	*AXL*	ENST00000301178	ENSG00000196155	*PLEKHG4*	ENST00000360461
ENSG00000186716	*BCR*	ENST00000305877	ENSG00000124126	*PREX1*	ENST00000371941
ENSG00000170312	*CDK1*	ENST00000395284	ENSG00000046889	*PREX2*	ENST00000288368
ENSG00000185324	*CDK10*	ENST00000353379	ENSG00000112655	*PTK7*	ENST00000481273
ENSG00000008128	*CDK11A*	ENST00000404249	ENSG00000196396	*PTPN1*	ENST00000371621
ENSG00000248333	*CDK11B*	ENST00000407249	ENSG00000179295	*PTPN11*	ENST00000351677
ENSG00000167258	*CDK12*	ENST00000447079	ENSG00000127947	*PTPN12*	ENST00000248594
ENSG00000065883	*CDK13*	ENST00000181839	ENSG00000163629	*PTPN13*	ENST00000436978
ENSG00000058091	*CDK14*	ENST00000265741	ENSG00000152104	*PTPN14*	ENST00000366956
ENSG00000138395	*CDK15*	ENST00000450471	ENSG00000072135	*PTPN18*	ENST00000175756
ENSG00000102225	*CDK16*	ENST00000276052	ENSG00000175354	*PTPN2*	ENST00000309660
ENSG00000059758	*CDK17*	ENST00000261211	ENSG00000126542	*PTPN20CP*	ENST00000506185
ENSG00000117266	*CDK18*	ENST00000506784	ENSG00000070778	*PTPN21*	ENST00000556564
ENSG00000155111	*CDK19*	ENST00000368911	ENSG00000134242	*PTPN22*	ENST00000359785
ENSG00000123374	*CDK2*	ENST00000266970	ENSG00000076201	*PTPN23*	ENST00000265562
ENSG00000156345	*CDK20*	ENST00000325303	ENSG00000070159	*PTPN3*	ENST00000374541
ENSG00000250506	*CDK3*	ENST00000425876	ENSG00000088179	*PTPN4*	ENST00000263708
ENSG00000135446	*CDK4*	ENST00000257904	ENSG00000110786	*PTPN5*	ENST00000358540
ENSG00000164885	*CDK5*	ENST00000485972	ENSG00000111679	*PTPN6*	ENST00000456013
ENSG00000105810	*CDK6*	ENST00000265734	ENSG00000143851	*PTPN7*	ENST00000309017
ENSG00000134058	*CDK7*	ENST00000256443	ENSG00000169410	*PTPN9*	ENST00000306726
ENSG00000132964	*CDK8*	ENST00000381527	ENSG00000006451	*RALA*	ENST00000005257

ENSG00000136807	*CDK9*	ENST00000373264	ENSG00000144118	*RALB*	ENST00000272519
ENSG00000100490	*CDKL1*	ENST00000395834	ENSG00000116473	*RAP1A*	ENST00000369709
ENSG00000138769	*CDKL2*	ENST00000429927	ENSG00000127314	*RAP1B*	ENST00000250559
ENSG00000006837	*CDKL3*	ENST00000265334	ENSG00000125249	*RAP2A*	ENST00000245304
ENSG00000205111	*CDKL4*	ENST00000378803	ENSG00000181467	*RAP2B*	ENST00000323534
ENSG00000008086	*CDKL5*	ENST00000379989	ENSG00000123728	*RAP2C*	ENST00000342983
ENSG00000182578	*CSF1R*	ENST00000286301	ENSG00000100302	*RASD2*	ENST00000216127
ENSG00000204580	*DDR1*	ENST00000376575	ENSG00000058335	*RASGRF1*	ENST00000419573
ENSG00000162733	*DDR2*	ENST00000367922	ENSG00000113319	*RASGRF2*	ENST00000265080
ENSG00000176490	*DIRAS1*	ENST00000323469	ENSG00000100276	*RASL10A*	ENST00000216101
ENSG00000165023	*DIRAS2*	ENST00000375765	ENSG00000141150	*RASL10B*	ENST00000268864
ENSG00000162595	*DIRAS3*	ENST00000370981	ENSG00000122035	*RASL11A*	ENST00000241463
ENSG00000107554	*DNMBP*	ENST00000324109	ENSG00000128045	*RASL11B*	ENST00000248706
ENSG00000114346	*ECT2*	ENST00000392692	ENSG00000103710	*RASL12*	ENST00000220062
ENSG00000203734	*ECT2L*	ENST00000423192	ENSG00000134533	*RERG*	ENST00000256953
ENSG00000146648	*EGFR*	ENST00000275493	ENSG00000111404	*RERGL*	ENST00000229002
ENSG00000187682	*ERAS*	ENST00000338270	ENSG00000165731	*RET*	ENST00000355710
ENSG00000141736	*ERBB2*	ENST00000269571	ENSG00000106615	*RHEB*	ENST00000262187
ENSG00000065361	*ERBB3*	ENST00000267101	ENSG00000167550	*RHEBL1*	ENST00000301068
ENSG00000178568	*ERBB4*	ENST00000342788	ENSG00000143622	*RIT1*	ENST00000368322
ENSG00000152767	*FARP1*	ENST00000319562	ENSG00000152214	*RIT2*	ENST00000326695
ENSG00000006607	*FARP2*	ENST00000264042	ENSG00000185483	*ROR1*	ENST00000371079
ENSG00000102302	*FGD1*	ENST00000375135	ENSG00000169071	*ROR2*	ENST00000375708
ENSG00000146192	*FGD2*	ENST00000274963	ENSG00000047936	*ROS1*	ENST00000368508
ENSG00000127084	*FGD3*	ENST00000375482	ENSG00000126458	*RRAS*	ENST00000246792
ENSG00000139132	*FGD4*	ENST00000427716	ENSG00000133818	*RRAS2*	ENST00000256196
ENSG00000154783	*FGD5*	ENST00000285046	ENSG00000163785	*RYK*	ENST00000296084
ENSG00000180263	*FGD6*	ENST00000343958	ENSG00000115904	*SOS1*	ENST00000426016
ENSG00000077782	*FGFR1*	ENST00000425967	ENSG00000100485	*SOS2*	ENST00000216373
ENSG00000066468	*FGFR2*	ENST00000457416	ENSG00000182957	*SPATA13*	ENST00000424834
ENSG00000068078	*FGFR3*	ENST00000340107	ENSG00000060140	*STYK1*	ENST00000075503
ENSG00000160867	*FGFR4*	ENST00000292408	ENSG00000120156	*TEK*	ENST00000380036
ENSG00000102755	*FLT1*	ENST00000282397	ENSG00000156299	*TIAM1*	ENST00000286827
ENSG00000122025	*FLT3*	ENST00000241453	ENSG00000146426	*TIAM2*	ENST00000461783
ENSG00000037280	*FLT4*	ENST00000261937	ENSG00000066056	*TIE1*	ENST00000372476
ENSG00000174775	*HRAS*	ENST00000451590	ENSG00000038382	*TRIO*	ENST00000344204
ENSG00000140443	*IGF1R*	ENST00000268035	ENSG00000092445	*TYRO3*	ENST00000263798
ENSG00000171105	*INSR*	ENST00000302850	ENSG00000141968	*VAV1*	ENST00000602142
ENSG00000027644	*INSRR*	ENST00000368195	ENSG00000160293	*VAV2*	ENST00000371850
ENSG00000205726	*ITSN1*	ENST00000381318	ENSG00000134215	*VAV3*	ENST00000370056

Table A10 - Genes with Gene Ontology terms indicating role in telomere function used in analysis

Gene identifier	Gene name	Canonical transcript	Gene identifier	Gene name	Canonical transcript
ENSG00000102977	*ACD*	ENST00000393919	ENSG00000039650	*PNKP*	ENST00000322344
ENSG00000100823	*APEX1*	ENST00000216714	ENSG00000014138	*POLA2*	ENST00000265465
ENSG00000149311	*ATM*	ENST00000278616	ENSG00000062822	*POLD1*	ENST00000440232
ENSG00000175054	*ATR*	ENST00000350721	ENSG00000106628	*POLD2*	ENST00000406581
ENSG00000178999	*AURKB*	ENST00000585124	ENSG00000077514	*POLD3*	ENST00000263681
ENSG00000105173	*CCNE1*	ENST00000262643	ENSG00000175482	*POLD4*	ENST00000312419
ENSG00000175305	*CCNE2*	ENST00000520509	ENSG00000177084	*POLE*	ENST00000320574
ENSG00000166226	*CCT2*	ENST00000299300	ENSG00000100479	*POLE2*	ENST00000216367
ENSG00000163468	*CCT3*	ENST00000295688	ENSG00000148229	*POLE3*	ENST00000374171
ENSG00000115484	*CCT4*	ENST00000394440	ENSG00000115350	*POLE4*	ENST00000483063
ENSG00000150753	*CCT5*	ENST00000280326	ENSG00000128513	*POT1*	ENST00000357628
ENSG00000146731	*CCT6A*	ENST00000275603	ENSG00000204569	*PPP1R10*	ENST00000376511
ENSG00000135624	*CCT7*	ENST00000258091	ENSG00000198056	*PRIM1*	ENST00000338193
ENSG00000156261	*CCT8*	ENST00000286788	ENSG00000146143	*PRIM2*	ENST00000607273
ENSG00000178971	*CTC1*	ENST00000315684	ENSG00000065675	*PRKCQ*	ENST00000263125
ENSG00000168036	*CTNNB1*	ENST00000349496	ENSG00000110958	*PTGES3*	ENST00000262033
ENSG00000118655	*DCLRE1B*	ENST00000369563	ENSG00000113522	*RAD50*	ENST00000265335
ENSG00000172795	*DCP2*	ENST00000389063	ENSG00000051180	*RAD51*	ENST00000382643
ENSG00000174953	*DHX36*	ENST00000496811	ENSG00000185379	*RAD51D*	ENST00000590016
ENSG00000138346	*DNA2*	ENST00000399180	ENSG00000160957	*RECQL4*	ENST00000428558
ENSG00000012061	*ERCC1*	ENST00000013807	ENSG00000035928	*RFC1*	ENST00000381897
ENSG00000175595	*ERCC4*	ENST00000311895	ENSG00000049541	*RFC2*	ENST00000055077
ENSG00000171824	*EXOSC10*	ENST00000376936	ENSG00000133119	*RFC3*	ENST00000380071
ENSG00000151876	*FBXO4*	ENST00000281623	ENSG00000163918	*RFC4*	ENST00000392481
ENSG00000168496	*FEN1*	ENST00000305885	ENSG00000111445	*RFC5*	ENST00000454402
ENSG00000109534	*GAR1*	ENST00000226796	ENSG00000080345	*RIF1*	ENST00000243326
ENSG00000163938	*GNL3*	ENST00000418458	ENSG00000132383	*RPA1*	ENST00000254719
ENSG00000166923	*GREM1*	ENST00000300177	ENSG00000117748	*RPA2*	ENST00000373912
ENSG00000083307	*GRHL2*	ENST00000251808	ENSG00000106399	*RPA3*	ENST00000223129
ENSG00000147421	*HMBOX1*	ENST00000397358	ENSG00000258366	*RTEL1*	ENST00000508582
ENSG00000135486	*HNRNPA1*	ENST00000340913	ENSG00000026036	*RTEL1-TNFRSF6B*	ENST00000482936
ENSG00000122566	*HNRNPA2B1*	ENST00000354667	ENSG00000077463	*SIRT6*	ENST00000337491
ENSG00000092199	*HNRNPC*	ENST00000320084	ENSG00000132207	*SLX1A*	ENST00000251303
ENSG00000138668	*HNRNPD*	ENST00000313899	ENSG00000188827	*SLX4*	ENST00000294008
ENSG00000153187	*HNRNPU*	ENST00000283179	ENSG00000157106	*SMG1*	ENST00000446231
ENSG00000080824	*HSP90AA1*	ENST00000334701	ENSG00000198952	*SMG5*	ENST00000361813
ENSG00000096384	*HSP90AB1*	ENST00000371554	ENSG00000070366	*SMG6*	ENST00000263073
ENSG00000004487	*KDM1A*	ENST00000400181	ENSG00000116698	*SMG7*	ENST00000507469
ENSG00000136826	*KLF4*	ENST00000374672	ENSG00000060688	*SNRNP40*	ENST00000263694
ENSG00000155858	*LSM11*	ENST00000286307	ENSG00000125835	*SNRPB*	ENST00000438552
ENSG00000076984	*MAP2K7*	ENST00000397979	ENSG00000100028	*SNRPD3*	ENST00000215829
ENSG00000085511	*MAP3K4*	ENST00000392142	ENSG00000182004	*SNRPE*	ENST00000414487
ENSG00000100030	*MAPK1*	ENST00000215832	ENSG00000067066	*SP100*	ENST00000340126
ENSG00000181085	*MAPK15*	ENST00000338033	ENSG00000197122	*SRC*	ENST00000373578
ENSG00000102882	*MAPK3*	ENST00000263025	ENSG00000120438	*TCP1*	ENST00000321394
ENSG00000089022	*MAPKAPK5*	ENST00000551404	ENSG00000100726	*TELO2*	ENST00000262319
ENSG00000020922	*MRE11A*	ENST00000323929	ENSG00000257949	*TEN1*	ENST00000397640
ENSG00000136997	*MYC*	ENST00000377970	ENSG00000129566	*TEP1*	ENST00000262715
ENSG00000139579	*NABP2*	ENST00000380198	ENSG00000147601	*TERF1*	ENST00000276603
ENSG00000145414	*NAF1*	ENST00000274054	ENSG00000132604	*TERF2*	ENST00000603068
ENSG00000135372	*NAT10*	ENST00000257829	ENSG00000166848	*TERF2IP*	ENST00000300086

ENSG00000104320	*NBN*	ENST00000265433	ENSG00000164362	*TERT*	ENST00000310581
ENSG00000115053	*NCL*	ENST00000322723	ENSG00000092330	*TINF2*	ENST00000267415
ENSG00000117650	*NEK2*	ENST00000366999	ENSG00000173273	*TNKS*	ENST00000310430
ENSG00000151414	*NEK7*	ENST00000367385	ENSG00000149115	*TNKS1BP1*	ENST00000532437
ENSG00000145912	*NHP2*	ENST00000274606	ENSG00000107854	*TNKS2*	ENST00000371627
ENSG00000182117	*NOP10*	ENST00000328848	ENSG00000067369	*TP53BP1*	ENST00000382044
ENSG00000143748	*NVL*	ENST00000281701	ENSG00000005007	*UPF1*	ENST00000262803
ENSG00000121274	*PAPD5*	ENST00000436909	ENSG00000151461	*UPF2*	ENST00000356352
ENSG00000169116	*PARM1*	ENST00000307428	ENSG00000169062	*UPF3A*	ENST00000375299
ENSG00000140694	*PARN*	ENST00000437198	ENSG00000141499	*WRAP53*	ENST00000316024
ENSG00000143799	*PARP1*	ENST00000366794	ENSG00000165392	*WRN*	ENST00000298139
ENSG00000041880	*PARP3*	ENST00000398755	ENSG00000079246	*XRCC5*	ENST00000392133
ENSG00000102699	*PARP4*	ENST00000381989	ENSG00000196419	*XRCC6*	ENST00000359308
ENSG00000132646	*PCNA*	ENST00000379160	ENSG00000114127	*XRN1*	ENST00000264951
ENSG00000140451	*PIF1*	ENST00000268043	ENSG00000204859	*ZBTB48*	ENST00000377674
ENSG00000254093	*PINX1*	ENST00000314787	ENSG00000138311	*ZNF365*	ENST00000410046
ENSG00000135549	*PKIB*	ENST00000258014	ENSG00000180532	*ZSCAN4*	ENST00000318203
ENSG00000140464	*PML*	ENST00000268058			

Appendix 4 - Tumour type labels designated as arising from GTEx tissues

Table A11 – Tumour type labels designated as arising from GTEx tissues

GTEx tisue	GTEx tissue filename	Single word equivalent 1	Single word equivalent 2	Single word equivalent 3	Single word equivalent 4
Adipose_Subcutaneous	Adipose_Subcutaneous.v7.signif_variant_gene_pairs.txt	Lipoma			
Adipose_Visceral_Omentum	Adipose_Visceral_Omentum.v7.signif_variant_gene_pairs.txt	No occurences			
Adrenal_Gland	Adrenal_Gland.v7.signif_variant_gene_pairs.txt	Phaeochromoctoma	ACC		
Artery_Aorta	Artery_Aorta.v7.signif_variant_gene_pairs.txt	No occurences			
Artery_Coronary	Artery_Coronary.v7.signif_variant_gene_pairs.txt	No occurences			
Artery_Tibial	Artery_Tibial.v7.signif_variant_gene_pairs.txt	No occurences			
Brain_Amygdala	Brain_Amygdala.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Anterior_cingulate_cortex_BA24	Brain_Anterior_cingulate_cortex_BA24.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Caudate_basal_ganglia	Brain_Caudate_basal_ganglia.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Cerebellar_Hemisphere	Brain_Cerebellar_Hemisphere.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Cerebellum	Brain_Cerebellum.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Cortex	Brain_Cortex.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Frontal_Cortex_BA9	Brain_Frontal_Cortex_BA9.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Hippocampus	Brain_Hippocampus.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Hypothalamus	Brain_Hypothalamus.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Nucleus_accumbens_basal_ganglia	Brain_Nucleus_accumbens_basal_ganglia.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Putamen_basal_ganglia	Brain_Putamen_basal_ganglia.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Spinal_cord_cervical_c-1	Brain_Spinal_cord_cervical_c-1.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Brain_Substantia_nigra	Brain_Substantia_nigra.v7.signif_variant_gene_pairs.txt	CNS	CNS nerve sheath		
Breast_Mammary_Tissue	Breast_Mammary_Tissue.v7.signif_variant_gene_pairs.txt	Breast			
Cells_EBV-transformed_lymphocytes	Cells_EBV-transformed_lymphocytes.v7.signif_variant_gene_pairs.txt	Haem. lymphoid			
Cells_Transformed_fibroblasts	Cells_Transformed_fibroblasts.v7.signif_variant_gene_pairs.txt	N/A			
Colon_Sigmoid	Colon_Sigmoid.v7.signif_variant_gene_pairs.txt	Colorectal			

Colon_Transverse	Colon_Transverse.v7.signif_variant_gene_pairs.txt	Colorectal			
Esophagus_Gastroesophageal_Junction	Esophagus_Gastroesophageal_Junction.v7.signif_variant_gene_pairs.txt	Oesophagus			
Esophagus_Mucosa	Esophagus_Mucosa.v7.signif_variant_gene_pairs.txt	Oesophagus			
Esophagus_Muscularis	Esophagus_Muscularis.v7.signif_variant_gene_pairs.txt	Oesophagus			
Heart_Atrial_Appendage	Heart_Atrial_Appendage.v7.signif_variant_gene_pairs.txt	Cardiac myxoma			
Heart_Left_Ventricle	Heart_Left_Ventricle.v7.signif_variant_gene_pairs.txt	No occurences			
Liver	Liver.v7.signif_variant_gene_pairs.txt	No occurences			
Lung	Lung.v7.signif_variant_gene_pairs.txt	Lung			
Minor_Salivary_Gland	Minor_Salivary_Gland.v7.signif_variant_gene_pairs.txt	Salivary gland			
Muscle_Skeletal	Muscle_Skeletal.v7.signif_variant_gene_pairs.txt	Soft tissue sarcoma			
Nerve_Tibial	Nerve_Tibial.v7.signif_variant_gene_pairs.txt	PNS nerve sheath benign	PNS nerve sheath	Nerve sheath benign	
Ovary	Ovary.v7.signif_variant_gene_pairs.txt	Ovary			
Pancreas	Pancreas.v7.signif_variant_gene_pairs.txt	Pancreas			
Pituitary	Pituitary.v7.signif_variant_gene_pairs.txt	Pituitary			
Prostate	Prostate.v7.signif_variant_gene_pairs.txt	Prostate			
Skin_Not_Sun_Exposed_Suprapubic	Skin_Not_Sun_Exposed_Suprapubic.v7.signif_variant_gene_pairs.txt	NMSC	Melanoma	Skin benign	
Skin_Sun_Exposed_Lower_leg	Skin_Sun_Exposed_Lower_leg.v7.signif_variant_gene_pairs.txt	NMSC	Melanoma	Skin benign	
Small_Intestine_Terminal_Ileum	Small_Intestine_Terminal_Ileum.v7.signif_variant_gene_pairs.txt	Small bowel	GINET		
Spleen	Spleen.v7.signif_variant_gene_pairs.txt	No occurences			
Stomach	Stomach.v7.signif_variant_gene_pairs.txt	Gastric			
Testis	Testis.v7.signif_variant_gene_pairs.txt	Testicular			
Thyroid	Thyroid.v7.signif_variant_gene_pairs.txt	Thyroid			
Uterus	Uterus.v7.signif_variant_gene_pairs.txt	Endometrial	Uterine leiomyoma	Uterine sarcoma	
Vagina	Vagina.v7.signif_variant_gene_pairs.txt	No occurences			
Whole_Blood	Whole_Blood.v7.signif_variant_gene_pairs.txt	Haem. lymphoid	Haem. myeloid	Haem. polycythaemia	Haem. thrombocythaemia

ACC - Adrenocortical carcinoma, CNS – Central nervous system, Haem. – Haematological, PNS – Peripheral nervous system

Appendix 5 - Detail and validation of structural variants called from whole genome sequencing data and described in Chapter 3 and Chapter 6

Variant 1 – Chromosome 17 deletion involving *FLCN*

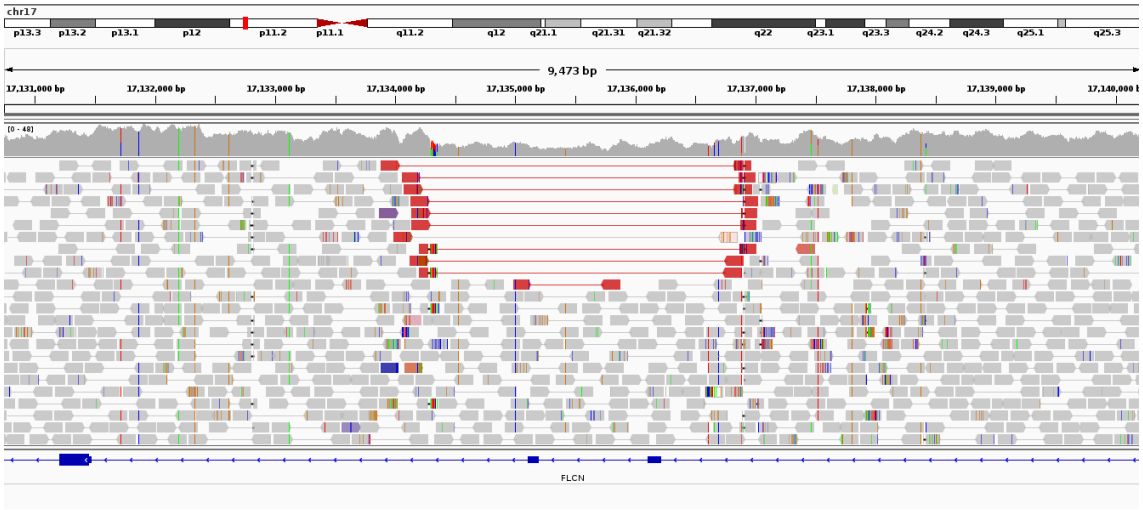
Coordinates – 17:17134310-17136696 (Manta), 17:17134474-17137867 (Canvas)

Description – Deletion of exon 2

Phenotype - Breast cancer, 46y; Pulmonary lymphangioleiomyomatosis, 47y

Sanger sequencing validation comment - Long Range PCR with primers to amplify across 17:17134310-17137867 shows wild type and deleted allele as two bands (wild type allele at ~5,700bp and deleted allele at ~3,500bp). Deletion confirmed though no sequence data from across breakpoints.

Figure A1 – IGV plot pertaining to chromosome 17 deletion involving *FLCN*



Reads viewed as pairs and grouped by insert size. Read pairs corresponding to deletion shown by large insert size (highlighted in red).

Variant 2 – Chromosome 10 inversion involving *PTEN*

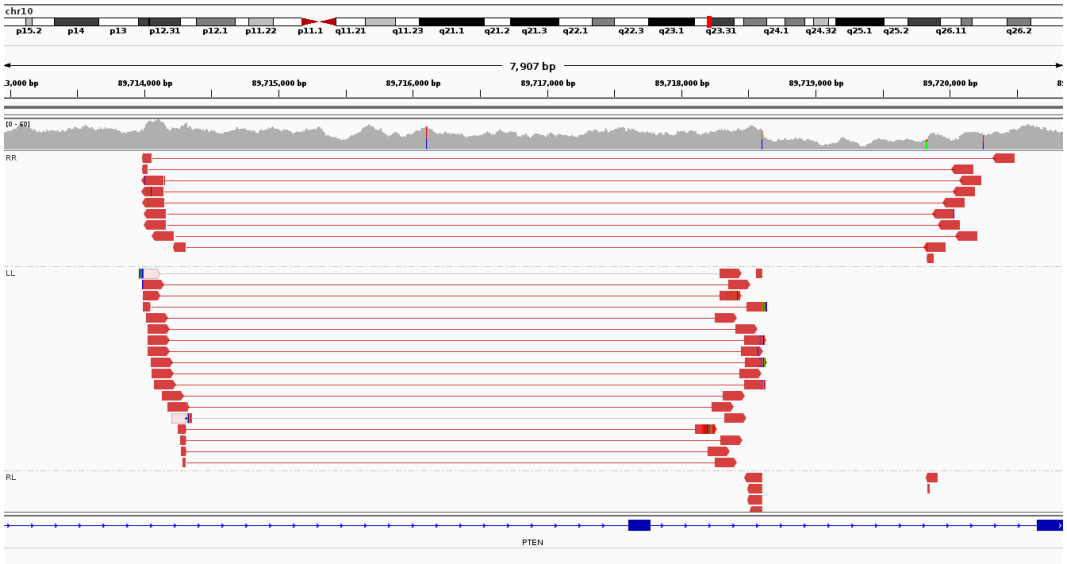
Coordinates – 2:89713996-89719837 (Manta)

Description – Inversion of exon 7

Phenotype – Breast cancer, 45y

Sanger sequencing validation comment - PCR primers across breakpoint 10:89719837 produce unique fragment. Sanger sequence data shows inversion is present.

Figure A2 - IGV plot pertaining to chromosome 10 inversion involving *PTEN*



Reads viewed as pairs and grouped by pair orientation. Read pairs corresponding to inversion shown by right-right (RR) or left-left (LL) orientation (highlighted in red).

Variant 3 – Chromosome 18:9 translocation involving *SMAD4*

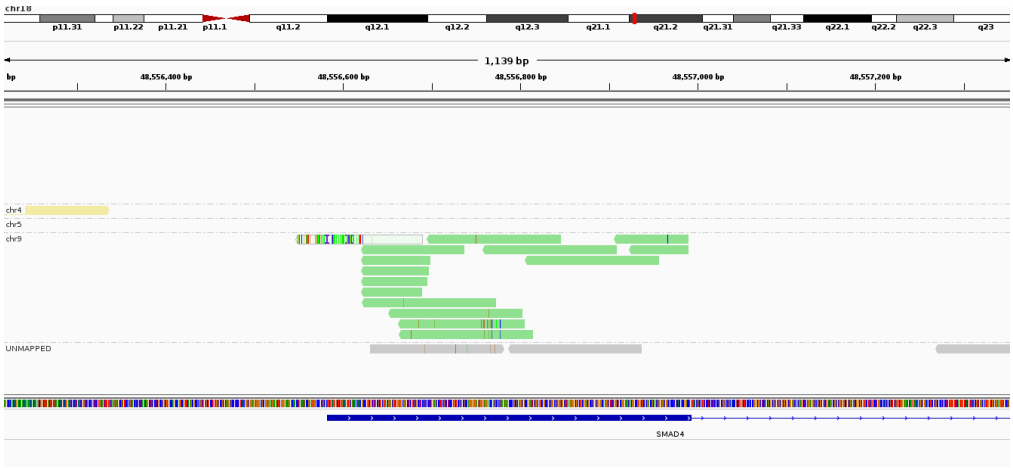
Coordinates – 18:48556624-9:127732713 (Manta)

Description – Translocation with breakpoint within untranslated part of exon 1

Phenotype - Central nervous system tumour, 45y

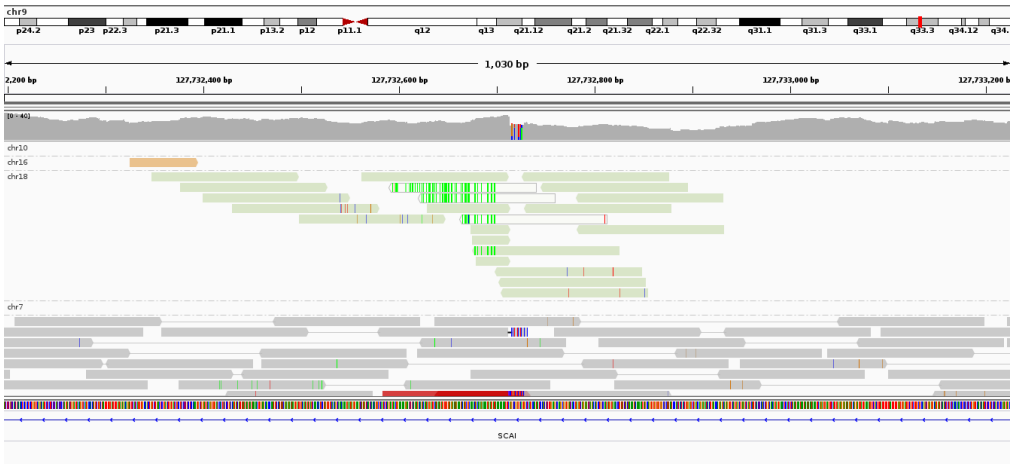
Sanger sequencing validation comment - ~700bp unique fragment with primers MP007R-MP008R across SMAD4 Translocation Breakpoint. Sanger sequencing of the unique fragments showed fragment maps to chromosome 9 at translocation breakpoint 9:127732713 and fusion of chr18 transcript into chr9. Translocation confirmed.

Figure A3.1 - IGV plot pertaining to chromosome 18:9 translocation involving *SMAD4* – Breakpoint at *SMAD4*



Reads grouped by alignment chromosome of mate pair. Read pairs corresponding to translocation shown by alignment to chromosome 5 (highlighted in green).

Figure A3.2 - IGV plot pertaining to chromosome 18:9 translocation involving *SMAD4* – Breakpoint at *SCAI*



Reads grouped by alignment chromosome of mate pair. Read pairs corresponding to translocation shown by alignment to chromosome 18 (highlighted in pale green).

Variant 4 – Chromosome 9 tandem duplication involving *TSC1*

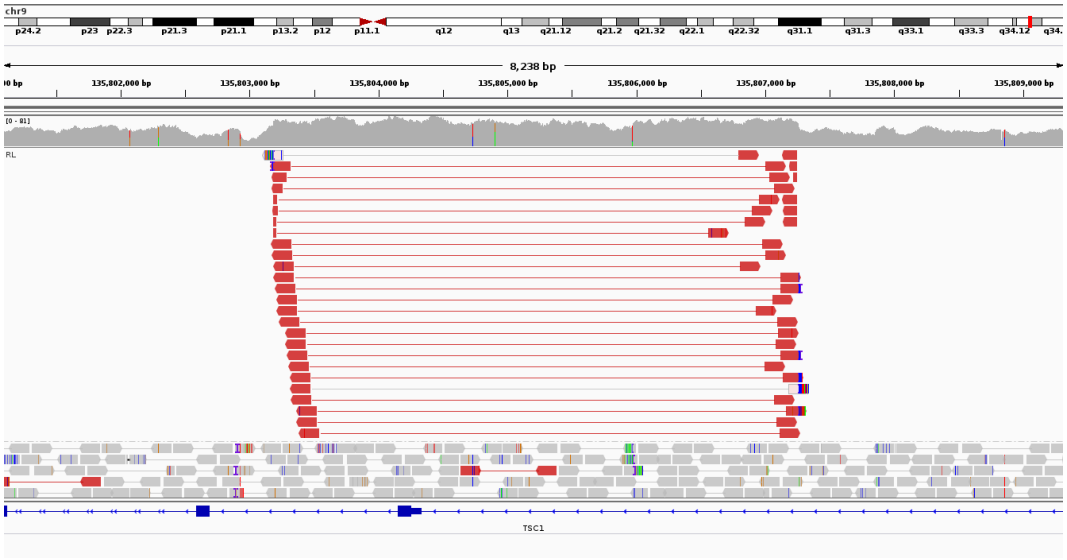
Coordinates – 9:135803187-135807261 (Manta)

Description – Duplication involving exon 3

Phenotype – Testicular cancer, 47y; Prostate cancer, 64y; Lung cancer, 70y

Sanger sequencing validation comment - Obtained unique fragment in that would only be amplified if tandem duplication present. Sanger sequencing of the fragment across breakpoint successful. Tandem duplication confirmed.

Figure A4 - IGV plot pertaining to chromosome 9 tandem duplication involving *TSC1*



Reads viewed as pairs and grouped by pair orientation. Read pairs corresponding to inversion shown by right-left (RL) orientation (highlighted in red).

Variant 5 – Chromosome 16 inversion involving *TSC2*

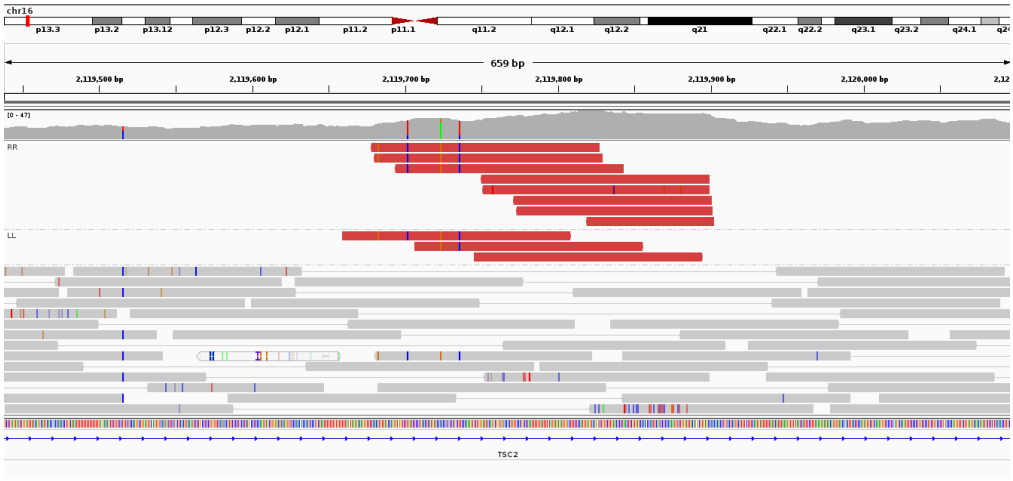
Coordinates – 16:1566500-2119769 (Manta)

Description – Inversion with breakpoint in intron 16-17

Phenotype - Small bowel cancer, 42y; Colorectal cancer, 43y

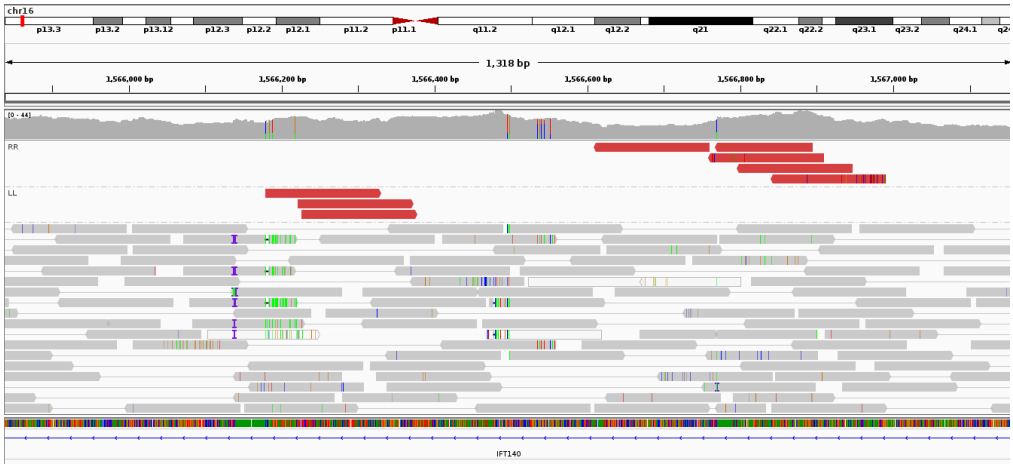
Sanger sequencing validation comment - PCR primers across breakpoint 16:1566500 gave two bands, the wild-type sized band and a slightly larger band. Gel purification and Sanger sequencing of the unique larger band demonstrates that the inversion is present. Unable to sequence to confirm at breakpoint 16:2119769. Inversion confirmed.

Figure A5.1 - IGV plot pertaining to chromosome 16 inversion involving *TSC2* – Breakpoint at *TSC2*



Reads grouped by pair orientation. Read pairs corresponding to inversion shown by right-right (RR) and left-left (LL) orientation (highlighted in red). Breakpoints of inversion too distant for viewing as read pairs.

Figure A5.2 - IGV plot pertaining to chromosome 16 inversion involving *TSC2* – Breakpoint at *IFT140*



Reads grouped by pair orientation. Read pairs corresponding to inversion shown by right-right (RR) and left-left (LL) orientation (highlighted in red). Breakpoints of inversion too distant for viewing as read pairs.

Variant 6 – Chromosome 1 deletion involving *FH*

Coordinates – 1: 237244834-242310908 (Canvas)
Description – Full gene deletion
Phenotype - Multiple cutaneous leiomyomata, <55y
Sanger sequencing validation comment - Long range PCR with primers to amplify across 1:237244834-242310908 gives unique ~7500bp fragment. Gel purification and attempt at Sanger sequencing. No data obtained for exact breakpoints. Deletion probably confirmed.

Variant 7 – Chromosome 17:10 translocation involving *FLCN*

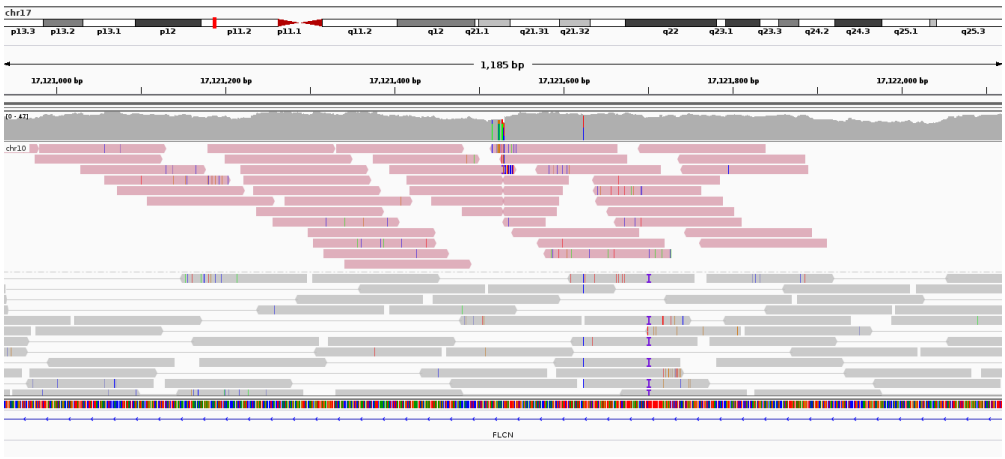
Coordinates – 17:17121531-10:43731507 (Manta)

Description – Translocation with breakpoint in intron 9-10

Phenotype - Multiple fibrofolliculomas, 18y; Renal cell carcinoma, 53y

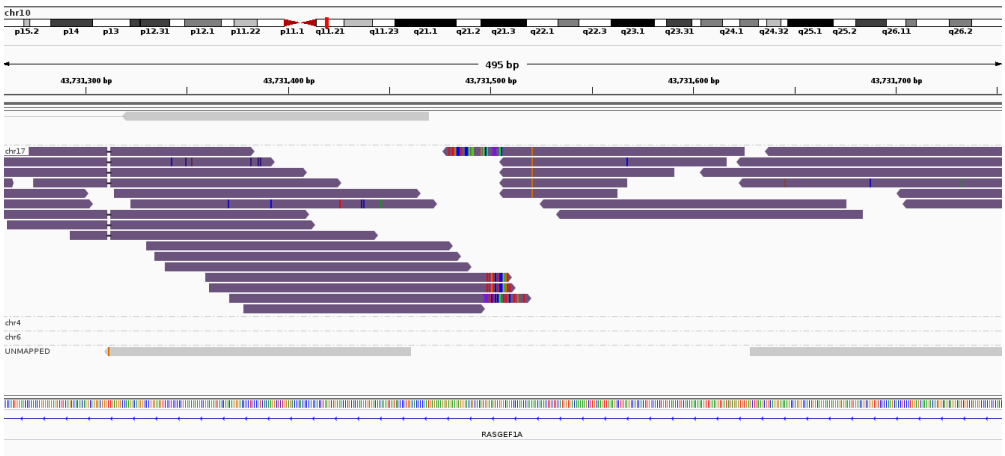
Sanger sequencing validation comment – Amplification demonstrated specific bands to confirm the translocation. Sanger sequencing data obtained from those amplicons putting breakpoints at ~17:17121526 and ~10:43731498.

Figure A6.1 - IGV plot pertaining to chromosome 17:10 translocation involving *FLCN* – Breakpoint at *FLCN*



Reads grouped by alignment chromosome of mate pair. Read pairs corresponding to translocation shown by alignment to chromosome 10 (highlighted in pink).

Figure A6.2 - IGV plot pertaining to chromosome 17:10 translocation involving *FLCN* – Breakpoint at *RASGEF1A*



Reads grouped by alignment chromosome of mate pair. Read pairs corresponding to translocation shown by alignment to chromosome 17 (highlighted in purple).

Variant 8 – Chromosome 10:6 translocation affecting *HABP2*

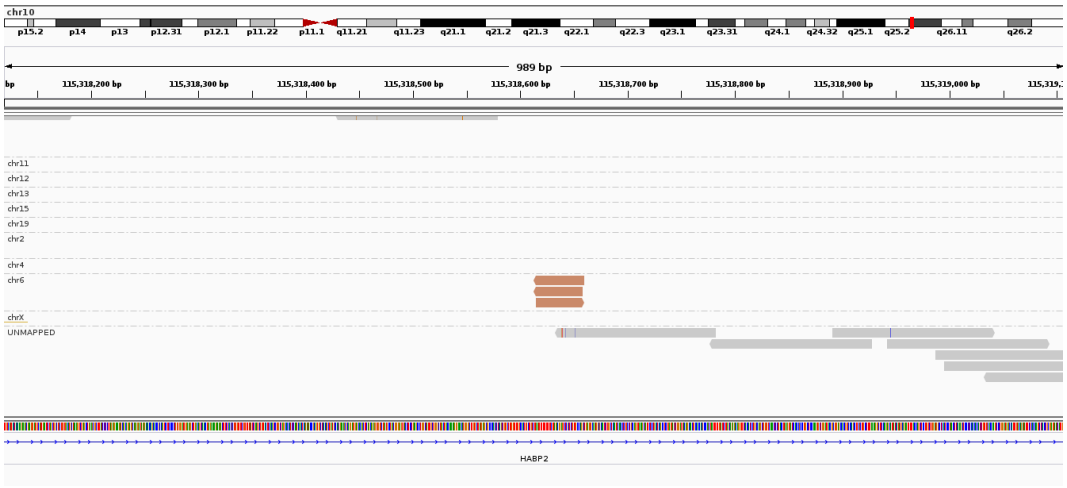
Coordinates – 10:115318616-6:7227789 (Manta)

Description – Translocation with breakpoint between exons 1 and 2 (both coding)

Phenotype - Breast cancer (bilateral), 46y; Colorectal cancer, 51y; Pancreatic cancer, 52y

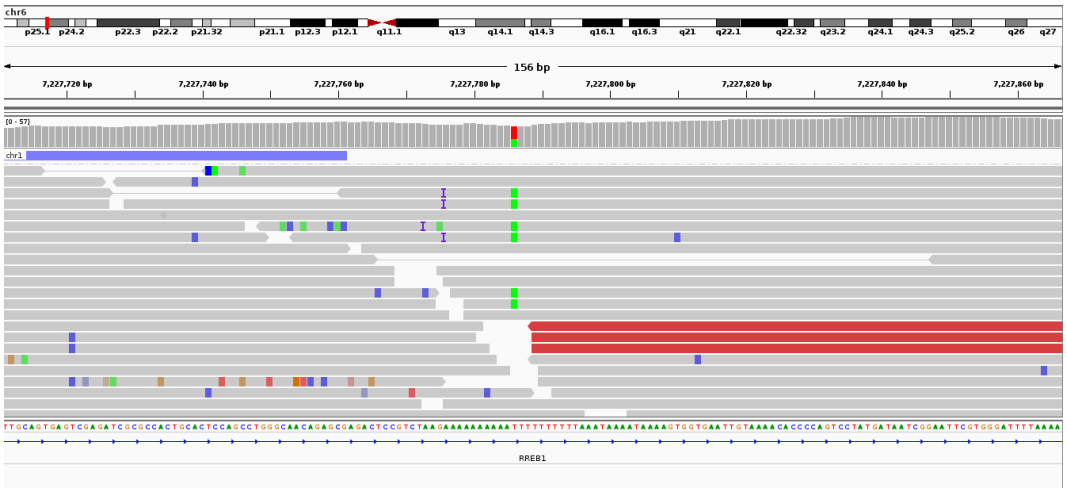
Sanger sequencing validation comment – Validation by Sanger sequencing not performed.

Figure A7.1 - IGV plot pertaining to chromosome 10:6 translocation affecting *HABP2* – Breakpoint at *HABP2*



Reads grouped by alignment chromosome of mate pair. Read pairs corresponding to translocation shown by alignment to chromosome 6 (highlighted in orange).

Figure A7.2 - IGV plot pertaining to chromosome 10:6 translocation affecting *HABP2* – Breakpoint at *RREB1*



Reads grouped by alignment chromosome of mate pair. No read pairs corresponding to translocation evident.

Variant 9 – Chromosome 10:5 deletion affecting *BMPR1A*

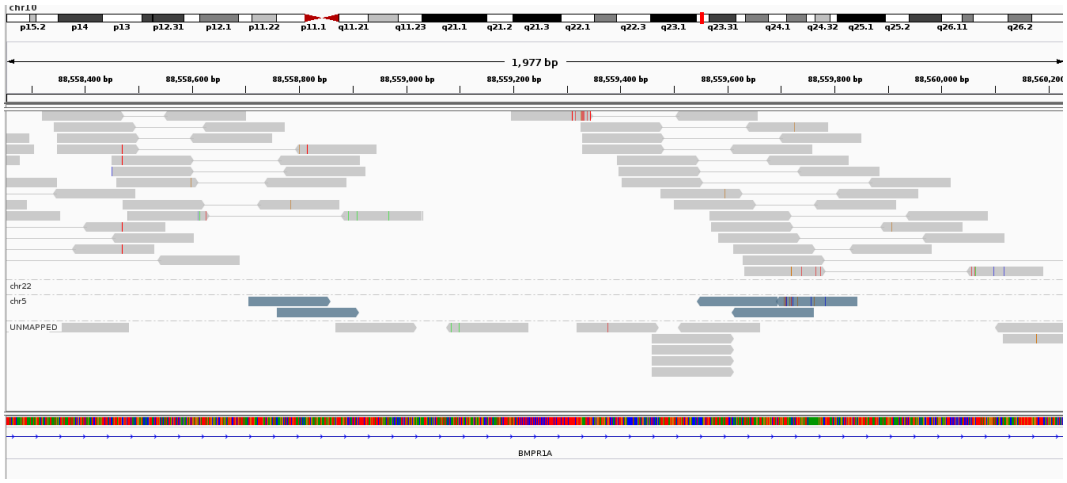
Coordinates – 10: 88559247-5:107163219 (Manta)

Description – Translocation with breakpoint between exons 1 and 2 (both non-coding)

Phenotype – Breast cancer, 52y; Central nervous system meningioma, 56y; Breast cancer, 58y; Aerodigestive tract cancer, 63y

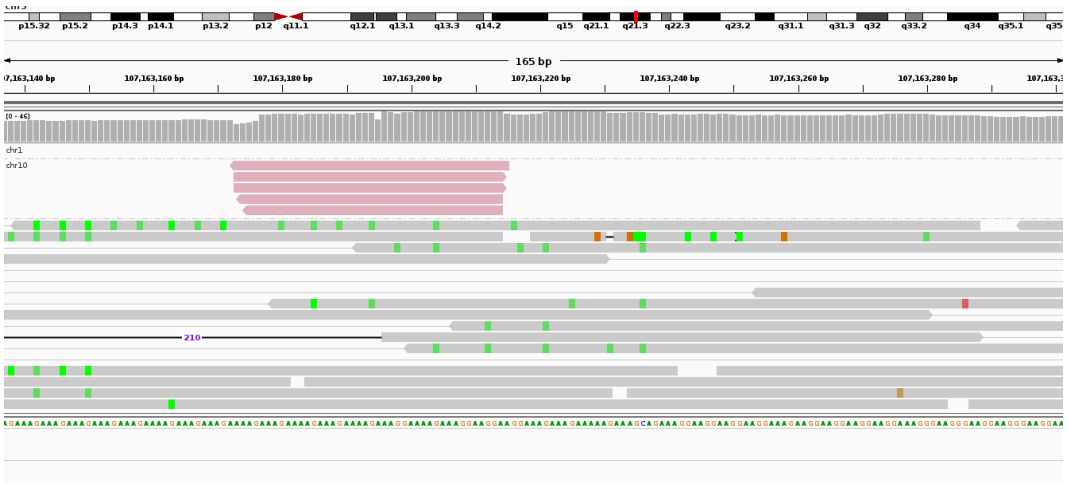
Sanger sequencing validation comment – Validation by Sanger sequencing not performed

Figure A8.1 - IGV plot pertaining to chromosome 10:5 translocation affecting *BMPR1A* – Breakpoint at *BMPR1A*



Reads grouped by alignment chromosome of mate pair. Read pairs corresponding to translocation shown by alignment to chromosome 5 (highlighted in blue).

Figure A8.2 - IGV plot pertaining to chromosome 10:5 translocation affecting *BMPR1A* – Breakpoint at 5q21.3



Reads grouped by alignment chromosome of mate pair. Read pairs corresponding to translocation shown by alignment to chromosome 10 (highlighted in pink).

Variant 10 – Chromosome 19 deletion affecting eQTL where variants reported to reduce *ZNF284* expression

Coordinates – 19: 43765327-43848192 (Canvas)

Description – Deletion of entire eQTL region

Phenotype – Prostate cancer, 54y; Colorectal cancer, 54y

Sanger sequencing validation comment – Validation by Sanger sequencing not performed